STATISTICS, STATISTICAL MODELLING AND DATA ANALYTICS

By:

Rajni
Assistant Professor
MAIT, Delhi

Ajay Kumar Gupta
Assistant Professor
MAIT, Delhi

Kanika Agarwal
Assistant Professor
MAIT, Delhi

SPECIMEN COPY
For..............................
NOT FOR SALE

S.K. KATARIA & SONS®
Educational Publishers
4885/109, Prakash Mahal, Dr. Subhash Bhargav Lane,
Opposite Delhi Medical Association, Daryaganj, New Delhi-110002 (INDIA)
Phone: +91-11-23243489, 43551243; *Mobile:* +91-9871775858
e-mail: katariabook@gmail.com; katariabooks@yahoo.com
Website: www.skkatariaandsons.com
(e-books, subscriptions and bulk buying also available on our website)

Published by:

S.K. Kataria & Sons®
4885/109, Prakash Mahal, Dr. Subhash Bhargav Lane,
Opposite Delhi Medical Association, Daryaganj, New Delhi-110002 (INDIA)
Phone: +91-11-23243489, 43551243; *Mobile:* +91-9871775858
e-mail: katariabook@gmail.com; katariabooks@yahoo.com
Website: www.skkatariaandsons.com

© ALL RIGHTS STRICTLY RESERVED

STATUTORY WARNING

Information contained in this book has been obtained by the author from sources believed to be reliable and is correct to the best of his knowledge. Every effort has been made to avoid errors and omissions and ensure accuracy. Any error or omission noted may be brought to the notice of the publisher which shall be taken care of in the forthcoming edition of this book. However, neither the publisher nor the author guarantees the accuracy or completeness of any information published herein, and neither the publisher nor the author takes any responsibility or liability for any inconvenience, expenses, losses or damage to anyone resulting from the contents of this book.

The book is meant for educational and learning purposes and there is no attempt on the part of publisher/author to render engineering and other professional services. If such services are required, the assistance of an appropriate professional should be sought.

The author of the book has taken all possible care to ensure that the contents of the book do not violate any existing copyright or other intellectual property rights of any person in any manner whatsoever. In the event, the author has been unable to track any source and if any copyright has been inadvertently infringed upon, the facts may be brought to the notice of the publisher in writing for corrective action.

Send all correspondence to : M/s. S.K. Kataria & Sons, New Delhi
Email: katariabook@gmail.com

First Edition: 2024

Laser Typesetting by :
S.K. Computers, Delhi-32

PRINTED AT :
AAPTE UDHYOG, NOIDA

Preface...

Welcome to the first edition of "Statistics, Statistical Modelling, and Data Analytics." This book is tailored for undergraduate students, with a particular emphasis on the curriculum of Guru Gobind Singh Indraprastha University (GGSIPU). However, the content is designed to be comprehensive and adaptable, making it suitable for students enrolled in other universities as well.

The field of statistics has evolved into a critical discipline that plays a fundamental role in various domains, from scientific research to business and industry. Understanding statistical concepts, methodologies, and data analytics is crucial for making informed decisions and solving real-world problems. This book is crafted to provide a solid foundation in these areas, catering to the specific needs and requirements of undergraduate students, including the following topics:

Statistics: Introduction & Descriptive Statistics, Data Visualization, Introduction to Probability Distributions, Hypothesis testing, Linear Algebra and Population Statistics, Mathematical Methods and Probability Theory, Sampling Distributions and Statistical Inference, Quantitative analysis.

Statistical Modelling: Linear models, regression analysis, analysis of variance, applications in various fields. Gauss-Markov theorem; geometry of least squares, subspace formulation of linear models, orthogonal projections; regression models, factorial experiments, analysis of covariance and model formulae; regression diagnostics, residuals, influence diagnostics, transformations, Box-Cox models, model selection and model building strategies, logistic regression models; Poisson regression models.

Data Analytics: Classes of open and closed set, concept of compactness, Metric spaces. concept of Cauchy sequence, completeness, compactness, and connectedness to solve the problems.

Advanced concepts in Data Analytics: Describe vector space, subspaces, independence of vectors, basis and dimension. Describe Eigen values, Eigen vectors and related results.

Rajni
Ajay Kumar Gupta
Kanika Agarwal

Acknowledgment...

The creation of this book is a collaborative effort, and we extend our sincere gratitude to our colleagues and students at Maharaja Agrasen Institute of Technology who played a crucial role in the creation and completion of this book. Their support, encouragement, and expertise have been invaluable.

Special thanks to our family and friends whose patience, understanding, and encouragement sustained us during the long hours of writing and editing.

Lastly, we extend our appreciation to the publisher S. K. Kataria & Sons who worked tirelessly with us to complete this project. Their collective efforts have left an indelible mark on the book.

Rajni
Ajay Kumar Gupta
Kanika Agarwal

CONTENT

1. Descriptive Statistics 1–15

1.1 Introduction 1
1.2 Descriptive Statistics 2
1.3 Descriptive Measures 3
    1.3.1 Measures of Centre 4
    1.3.2 Measures of Dispersion 6
1.4 Graphical Representation 11
    1.4.1 Pie Charts 11
    1.4.2 Bar Charts and Histograms .. 12
    1.4.3 Boxplots 13
Problems 14

2. Probability Theory and Probability Distribution 16–47

2.1 Introduction 16
2.2 Types of Random Variables 17
    2.2.1 Discrete Random Variables .. 18
    2.2.2 Continuous Random Variables 21
2.3 Jointly Distributed Random Variables 21
    2.3.1 Joint Distribution Function and its Properties 22
    2.3.2 Joint Discrete Distribution Function 23
    2.3.3 Joint Continuous Distribution Function 25
2.4 Independent and Dependent Random Variables 26
2.5 Expectation of the Sum of the Random Variables 26
2.6 Expectation of the Product of the Random Variables 27
2.7 Probability Distribution 29
    2.7.1 Binomial Distribution 29
    2.7.2 Poisson Distribution 37
    2.7.3 Poisson Process 40
    2.7.4 Normal Distribution (N.Dn) .. 42
Problems 46

3. Sampling Distribution 48–74

3.1 Introduction 48
3.2 Some Common Terms used in Sampling Distributions 49
3.3 Sampling Distributions 51
    3.3.1 The Sampling Distribution of the Mean (σ known) 52
    3.3.2 Central Limit Theorem 53
    3.3.3 Sampling Distribution of Proportions 54
    3.3.4 Sampling Distributions of Differences and Sums 54
    3.3.5 Sampling Distribution of Mean σ Unknown: *t*-distribution 55
    3.3.6 χ^2 – Distribution 56
    3.3.7 Sampling Distribution of Variance σ^2 58
    3.3.8 F-Distribution (Sampling Distribution of the Ratio of Two Sample Variances) 59
Problems 73

4. Confidence Intervals 75–84

4.1 Point Estimates 75
4.2 Confidence Intervals about the Mean (μ) when the Population Standard Deviation (σ) is Known 77
4.3 Confidence Intervals about the Mean (μ) when the Population Standard Deviation (σ) is Unknown .. 79
4.4 Confidence Intervals about the Population Proportion (p) 81
4.5 Confidence Interval Summary 82
Problems 83

5. Hypothesis Testing 85–135

5.1 The Fundamentals of Hypothesis Testing 85

5.1.1 Components of a Formal Hypothesis Test 86
5.1.2 The Null and Alternative Hypotheses 87
5.1.3 A Two-sided Test 87
5.1.4 A Right-sided Test 88
5.1.5 A Left-sided Test 88
5.1.6 Statistically Significant 89
5.1.7 Types of Errors 89
5.1.8 Power of the Test 90
5.2 Hypothesis Test about the Population Mean (μ) when the Population Standard Deviation (σ) is Known 91
5.2.1 Testing a Hypothesis Using Classical Approach 91
5.2.2 Testing a Hypothesis using *P*-values 96
5.3 Hypothesis Test about the Population Mean (μ) when the Population Standard Deviation (σ) is Unknown 100
5.3.1 Testing a Hypothesis Using Classical Approach 101
5.3.2 *P*-value Approach 105
5.4 Hypothesis Test for a Population Proportion (p) 106
5.5 Hypothesis Test about a Variance ... 110
5.5.1 One-sample Test for Testing the Hypotheses 111
5.6 Inferences about the Differences of Two Populations 112
5.6.1 Inferences about Two Means with Independent Samples (Assuming Unequal Variances) 112
5.6.2 Construct and Interpret a Confidence Interval About the Difference of Two Independent Means 115
5.6.3 P-value Approach 117
5.6.4 Pooled Two-sampled *t*-test (Assuming Equal Variances) 118
5.6.5 Inferences about Two Means with Dependent Samples–Matched Pairs 120
5.6.6 *P*-value Approach 123
5.6.7 Construct and Interpret a Confidence Interval about the Differences of the Data for Matched Pairs 124
5.6.8. Confidence Interval Interpretations 124
5.7 Inferences about Two Population Proportions 126
5.7.1 Using the *P*-Value Approach 129
5.8 F-Test for Comparing Two Population Variances 131
***Problems* 134**

6. Linear Algebra 136–152

6.1 Linear Equations 136
6.2 What is a System of Linear Equations? 136
6.3 Solving Linear Systems 140
6.4 Geometric Interpretation of the Solution set 143
6.5 Row Reduction and Echelon Forms 144
6.6 Reduced row echelon form (RREF) 147
6.7 Existence and Uniqueness of Solutions 151
***Problems* 151**

7. Simple Linear Regression ... 153–163

7.1 Introduction 153
7.2 Simple Linear Regression Model 153
7.3 Estimation of the Regression Parameters (β_0, β_1) 155
7.4 Meaning of Regression Parameters... 156

7.5 Properties of the Least Squares Estimators ... 158
7.6 Coefficient of Determination ... 160
Problems ... ***161***

8. Multiple Linear Regression Model ... 164–178

8.1 Introduction ... 164
8.2 Multiple Linear Regression Model ... 164
8.3 Meaning of Regression Parameter ... 166
8.4 Least-Squares Estimator for β ... 166
8.5 Properties of the Estimated Parameters ... 170
  8.5.1 Unbiases of $(\hat{\beta})$ ... 171
  8.5.2 Covar $(\hat{\beta}) = (X^TX)^{-1}\sigma^2$ ... 171
  8.5.3 Gauss-Markov Theorem ... 173
8.6 Geometry of Least Squares ... 174
  8.6.1 Subspace Formulation of Linear Models ... 175
  8.6.2 Orthogonal Projections ... 176
Problems ... ***177***

9. Analysis of Variance (ANOVA) Models ... 179–209

9.1 Introduction ... 179
9.2 Single Factor ANOVA Model ... 180
9.3 Analysis of Single Factor ANOVA ... 183
  9.3.1 Assumptions Required for the Analysis of Single Factor ANOVA ... 184
  9.3.2 Sums of Squares ... 185
  9.3.3 Testing of Hypotheses (Overall F-test for the Equality of the Factor Level Means) ... 185
9.4 Two-Factor ANOVA Model ... 188
9.5 Factorial Experiment ... 192
  9.5.1 Factorial Experiments with Factors at Two Levels (2^2 Factorial Experiment) ... 194
9.6 The Two-way Factorial Analysis ... 197
  9.6.1 The Linear Model for Two-way Factorial Experiments ... 197
  9.6.2 ANOVA for a Two-way Factorial Design ... 198
9.7 Analysis of Covariance ... 203
  9.7.1 The Analysis of Covariance (ANCOVA) Model ... 206
  9.7.2 One-Way Model with One Covariate ... 206
  9.7.3 Testing Hypotheses ... 207
Problems ... ***208***

10. Model Diagnostic ... 210–254

10.1 Introduction ... 210
10.2 Checking of the Linear Relationship Between Study and Explanatory Variables ... 211
10.3 Residual Analysis ... 213
10.4 Methods for Scaling Residuals ... 214
10.5 Diagnostic Plots ... 217
  10.5.1 Normal Probability Plot ... 218
  10.5.2 Plots of Residuals Against the Fitted Value ... 222
  10.5.3 Plots of Residuals Against Explanatory Variable ... 225
10.6 Transformation and Weighing to Correct Model Inadequacies ... 225
  10.6.1 Variance Stabilizing Transformations ... 225
  10.6.2 Transformations to Linearize the Model ... 227
10.7 Analytical Methods for Selecting a Transformation on Study Variable (The Box-Cox Method) ... 230
10.8 Leverage and Influence Diagnostics ... 234
  10.8.1 Leverage ... 235
  10.8.2 Measures of Influence ... 236
10.9 Variable Selection ... 239
  10.9.1 Omission/Exclusion of Relevant Variables ... 240
  10.9.2 Inclusion of Irrelevant Variables ... 242
10.10 Model Building ... 244
  10.10.1 Coefficient of Determination ... 244

10.10.2 Coefficient of Determination 245
10.10.3 Residual Mean Square 245
10.10.4 Mallow's C_p Statistics 246
10.10.5 Akaike's Information Criterion (AIC) 248
10.10.6 Bayesian Information Criterion (BIC) 249
10.10.7 PRESS Statistic 249
10.10.8 Partial F-statistics 249
10.11 Computational Techniques for Variable Selection 251
10.11.1 Use all Possible Explanatory Variables 251
10.11.2 Stepwise Regression Techniques 251
Problems 254

11. Logistic Regression Model and Poisson Regression Model 255–263

11.1 Logistic Regression Model 255
11.2 Linear Predictor and Link Functions 257
11.3 Maximum Likelihood Estimation of Parameters 258
11.4 Hypothesis Testing of Logistic Regression Model 259
11.5 Poisson Regression Model 260
11.6 Linear predictor and link functions: 261
11.7 Maximum Likelihood Estimation of Parameters 262
11.8 Hypothesis Testing of Poisson Regression Model 262
Problems 253

12. Data Analytics 264–278

12.1 Real analysis and Data Science 264
12.2 Basic Set Theory 265
12.3 Norms and Metric 267
12.3.1 Components of a Metric Space 268
12.3.2 Importance of Metric Space 269
12.4 Open set and Closed Set 269
12.4.1 Open Set 269
12.4.2 Closed Set 269
12.4.3 Unions of Open Sets, Intersections of Closed Sets 270
12.4.4 Example of Open and Closed Set 270
12.4.5 Example of Half-Open Interval 270
12.5 Sequences of Numbers 271
12.5.1 Sequences 271
12.5.2 Convergence 271
12.5.3 Cauchy Sequence 272
12.6 Compact Sets 272
12.6.1 Sequential Compactness .. 273
12.6.2 Key Properties of Compact Sets 273
12.6.3 Examples of Compact Sets 274
12.6.4 Importance of Compact Sets 274
12.7 Completeness in metric spaces 274
12.7.1 Examples of Complete Metric Space 275
12.8 Connected Metric Space 276
12.8.1 Theorems Based on Connectedness of Metric Space 276
12.8.2 Few Examples of Connected Metric Spaces 277
12.8.3 Problems on Connected Metric Space 277
Problems 278

13. Vector and Vector Space ... 279–293

13.1 Vector 279
13.2 Vector Addition and Scaler Multiplication 280
13.3 Hyperplanes and Lines in $\mathbb{R}^n$ 281
13.4 Directions and Magnitudes 283
13.5 Field 286
13.6 Vector Space 287
13.7 Spanning Sets 287
13.8 Subspaces 288

13.9 Linear Dependence and Independence 291
13.10 Basis and Dimensions 291
Problems *293*

14. Eigen Values and Eigen Vectors 294–305

14.1 Introduction 294
14.2 Cayley-Hamilton theorem, Characteristic Polynomial 295
14.3 Eigen Values and Eigen Vectors 296
14.4 Some other Properties of Eigen values 301
14.5 Diagonalization of a Matrix: 302
Problems ***304***

Appendix ***307–318***
Bibliography ***319***
Model Question Paper ***321–323***
Index ***325–328***

CHAPTER

1

DESCRIPTIVE STATISTICS

Chapter Outline

• Introduction • Descriptive Statistics • Descriptive Measures • Graphical Representation

1.1 INTRODUCTION

Statistics has evolved as a universal language of the sciences, and as an art of data analysis which can yield influential outcomes. In the field of natural resource management, which includes scientists, researchers, and administrators, we rely extensively on statistical analysis to address inquiries that emerge within the populations under our supervision. To illustrate:

- Has there been a substantial alteration in the average sawtimber volume within the red pine stands?
- Is there any evidence of a rise in the prevalence of invasive species within the Great Lakes region?
- What portion of the white-tailed deer in New Hampshire exhibits weights below the threshold considered healthy?
- Did fertilizers A, B, or C produce any discernible effects or corn yields?

However, before exploring diverse statistical methodologies, a brief overview of descriptive statistics is indispensable. These questions typically require statistical analysis for resolution. To tackle these inquiries effectively, it is crucial to acquire a well-constructed random sample from the relevent population. Once obtained we utilize descriptive statistics to organize and summarize our sample data. Subsequently, the next phase involves inferential statistics, enabling us to leverage our sample statistics and extend our findings to the entire population while assessing the reliability of the results. However, prior to delving into various statistical methodologies, a concise review of descriptive statistics is essential.

Statistics is the scientific field focussed on gathering, structuring, summarizing, analysing, and making sense of information.

Statistics is a branch of mathematics that involves the collection, analysis, interpretation, presentation, and organization of data. It provides methods for making inferences and decisions in the presence of uncertainty. There are two main branches of statistics:

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics: Descriptive Statistics encompasses the concise summarization and systematic organization of data to enhance its comprehensibility. Key metrics employed in descriptive statistics include the mean, median, mode, range, and standard deviation.

Inferential Statistics: Inferential statistics encompasses the utilization of data from a subset to draw inferences or predictions about a broader population. It encompasses methodologies such as hypothesis testing, confidence intervals, and regression analysis.

Statistics is widely applied in various fields, including science, business, economics, social sciences, and government, to draw meaningful conclusions from data and inform decision-making processes.

In this chapter, we will focus on descriptive statistics.

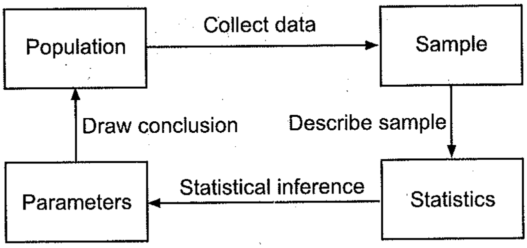


Fig. 1.1 Using sample statistics to estimate population parameters

1.2 DESCRIPTIVE STATISTICS

A population represents the subject of study, comprising all the elements within it. Population data entails the complete collection of these elements. For instance:

- Every single fish inhabiting the Long Lake.
- The entirety of lakes situated in the Adirondack Park.
- The total grizzly bear population residing in Yellowstone National Park.

In contrast, a sample refers to a portion of data selected from the population under investigation. For example:

- 100 fish selected at random from Long Lake.
- 25 lakes chosen through random selection within the Adirondack Park.
- A sample consisting of 60 grizzly bears, each with a designated home range in Yellowstone National Park.

Populations are characterized by descriptive metrics known as parameters. Inferences regarding these parameters rely on sample statistics. As an illustration, the population mean (μ) is approximated using the sample mean $(\bar{x})$, while the population variance (σ^2) finds its estimation in the sample variance (s^2).

Variables represent the characteristics of interest, such as:

- The length of fish dwelling in Long Lake.
- The pH levels of lakes in the Adirondack Park.
- The weights of grizzly bears residing in Yellowstone National Park.

Variables can be classified into two primary cateories: qualitative and quantitative. Qualitative variables encompass values that denote attributes or categories for, which mathematical operations are not applicable. Examples include gender, race, and petal colour. In contrast, quantitative variables consist of numeric values, such as measurements, to which mathematical operations can be applied. Illustrations of quantitative variables include age, height, and length.

Quantitative variables can be further classified into discrete and continuous variables. Discrete variables involve a finite or countable set of potential values much like a collection of 'hens' capable of laying specific numbers of eggs, such as 1, 2, 13, and so on. The range of conceivable values for discrete variables is finite and well-defined.

Continuous variables have an infinite number of possible values. Think of ontinuous variables as 'cows'. Cows can give 4.6713245 gallons of milk, or 7.0918754 gallons of milk, or 13.272698 gallons of milk... There are almost an infinite number of values that a continuous variable could take on.

Problem 1.1 Is the variable qualitative or quantitative?

Species	Weight	Diamet	Zip Code
(qualitative	quantitative,	quantitative,	qualitative)

1.3 DESCRIPTIVE MEASURES

Descriptive characteristics that pertain to populations are referred to as parameters and are commonly denoted by Greek letters. The population mean is symbolized by μ (pronounced as '*mu*'), the population variance is represented as σ^2 (pronounced as 'sigma squared'), and the population standard deviation is expressed as σ (pronounced as 'sigma').

Conversely, descriptive properties pertaining to samples are termed as statistics and are typically expressed using Roman letters. The sample mean is denoted as $(\bar{x})$ (pronounced as 'x-bar'), the sample variance is indicated as s^2, and the sample standard deviation is represented by s. Sample statistics serve the purpose of estimating unknown population parameters.

In this section, we will delve into descriptive statistics, specifically focusing on measures of central tendency and measures of data variability. These descriptive statistics aid us in discerning both the central tendency and the extent of variability within the dataset.

1.3.1 Measures of Centre

Mean

The arithmetic mean of a variable, often referred to as the average, is calculated by summing up all the values and then dividing the total by the count of values.

The population mean is symbolized using the Greek letter μ (pronounced as 'mu'), while the sample mean is represented as $\bar{x}$ (pronounced as 'x-bar'). Generally, the sample mean serves as the most reliable and unbiased estimate of the population mean. Nevertheless, it can be influenced by extreme values, also known as outliers, and may not be the most suitable measure of central tendency when dealing with highly skewed data. The following formulas illustrate the computation of the population mean and the sample mean:

$$\text{Population Mean } (\mu)\text{: } \mu = \sum \frac{x_i}{N}$$

$$\text{Sample Mean } (\bar{x})\text{: } \bar{x} = \sum \frac{x_i}{n}$$

Here, x_i represents an individual element within the dataset, N signifies the total number of elements in the population, and n stands for the number of elements in the sample dataset.

Problem 1.2 Find the mean for the following sample data set: 6.4, 5.2, 7.9, 3.4.

$$\bar{x} = \frac{6.4+5.2+7.9+3.4}{4}$$

Median

The median of a variable is determined by identifying the middle value within a dataset when the data are arranged in an ascending order. It effectively divides the data into two equal halves, with 50% of the data points falling below the median and the remaining 50% above it. Importantly, the median is not significantly affected by the presence of outliers and can serve as a more suitable measure of central tendency, especially when dealing with highly skewed data.

The method for calculating the median varies based on the number of observations in the dataset.

To compute the median for a dataset with an odd number of values (where '*n*' is an odd number), you should first arrange the data in ascending order.

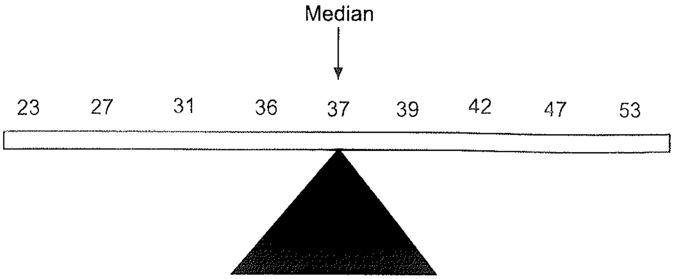


Fig. 1.2

The method of calculating the median varies based on the number of observations in the dataset.

To compute the median for a dataset with an odd number of values (where 'n' is an odd number), you should first arrange the data in ascending order.

Problem 1.3 Find the median for the following sample dataset:

$$23, 27, 29, 31, 35, 39, 40, 42, 44, 47, 51$$

The median is 39. It is the middle value that separates the lower 50% of the data from the upper 50% of the data.

To calculate the median with an even number of values (n is even), first sort the data from smallest to largest and take the average of the two middle values.

Problem 1.4 Find the median for the following sample data set:

$$23, 27, 29, 31, 35, 39, 40, 42, 44, 47$$

$$M = \frac{35 + 39}{2}$$

Mode

Mode, in statistics, refers to the most frequently occurring value in a dataset. It is particularly useful when working with qualitative data, where the values are categorical and lack numerical properties for mathematical operations such as addition, subtraction, multiplication, or division. As a result, computing the mean and median for categorical data is impractical. However, mode serves as a suitable measure of central tendency for such data sets. In contrast, dealing with quantitative data, mode is less commonly employed as a central measure, particularly when each value occurs only once, rendering the mode irrelevant.

Understanding the interplay between the mean and median is essential as it provides insights into the distribution characteristics of a variable. For instance, in a positively skewed distribution (or skewed right), the mean tends to increase due to a few larger observations that pull the distribution towards the right. Meanwhile, the median remains less affected by these extreme larger values. Consequently, in such a scenario, the mean will be greater than the median. In a symmetric distribution, the values of mean, median, and mode align closely. Conversely, in a negatively skewed distribution (or skewed left), the mean tends to decrease to account for a few smaller observations that pull the distribution towards the left. Once again, the median is less influenced by these extreme smaller values, resulting in the mean being lower than the median.

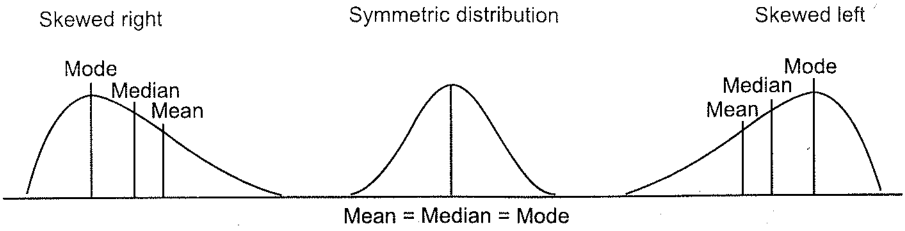


Fig. 1.3 Illustration of skewed and symmetric distributions

1.3.2 Measures of Dispersion

Measures of central tendency focus on the average or middle values within a dataset, while measures of dispersion analyze the extent or range of data variability. Variability indicales the degree to which values deviate from one another. Data points that are in close proximity to each other exhibit lower measures of variability, whereas values that are more widely dispersed display higher measures of variability.

Consider the two histograms provided below. Both sets possess an identical mean weight, yet the data points in Group A exhibit greater dispersion compared to those in Group B. Both groups exhibit an average weight of 267 pounds, but the weights in Group A show a higher degree of variation.

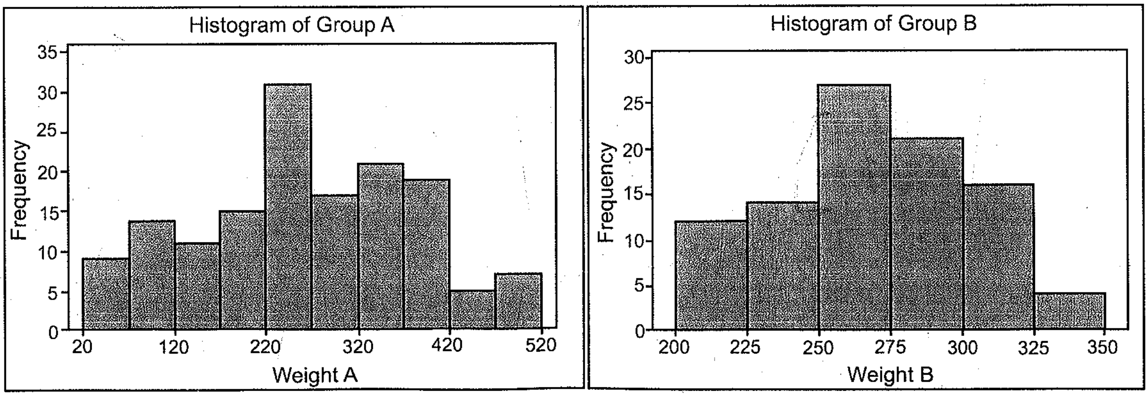


Fig. 1.4 Histograms of Group A and Group B

This section will examine five measures of dispersion: range, variance, standard deviation, standard error, and coefficient of variation.

Range

The range of a variable is determined by subtracting the smallest value from the largest value within a quantitative dataset, making it the most basic measure that relies solely on these two extreme values.

Problem 1.5 Find the range for the given dataset.

$$12, 29, 32, 34, 38, 49, 57$$

$$\text{Range} = 57 - 12 = 45$$

Variance

Variance is calculated by determining the squared differences between each value and the arithmetic mean. This method considers both positive and negative deviations. The sample variance (s^2) serves as an unbiased estimator of the population variance (σ^2), with ($n - 1$) degrees of freedom.

Degrees of freedom can be generally defined as the number of values involved minus the number of parameters estimated on the way to obtaining the particular estimate.

The sample variance is considered unbiased because of the adjustment in the denominator. If we used 'n' instead of '$n - 1$' in the denominator, we would consistently underestimate the actual population variance. To rectify this bias, we modify the denominator to '$n - 1$'.

Population Variance

$$\sigma^2 = \sum \frac{(x_i - \mu)^2}{N}$$

Sample Variance

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1} = \frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}{n-1}$$

Problem 1.6 Compute the variance of the sample data: 3, 5, 7. The sample mean is 5.

$$s^2 = \frac{(3-5)^2 + (5-5)^2 + (7-5)^2}{3-1} = 4$$

Standard Deviation

The standard deviation is the square root of the variance, and is applicable to both population and sample data. Although the sample variance serves as a positive and unbiased estimator for population variance, it is important to note that variance units are squared. The standard deviation is a frequently used numerical method that describes the spread of a variable's distribution. Specifically, the population standard deviation is denoted as σ (sigma), while the sample standard deviation is represented by s.

Population standard deviation

$$\sigma = \sqrt{\sigma^2}$$

Sample standard deviation

$$s = \sqrt{s^2}$$

Problem 1.7 Compute the standard deviation of the sample data: 3, 5, 7 with a sample mean of 5.

$$s = \sqrt{\frac{(3-5)^2+(5-5)^2+(7-5)^2}{3-1}} = \sqrt{4} = 2$$

Standard Error of the Means

Frequently, we utilize the sample mean, denoted as $\bar{x}$, to estimate the population mean, μ. For instance, if we aim to estimate the heights of eighty-year-old cherry trees, we can follow these steps:

- Randomly select 100 trees.
- Calculate the sample mean of these 100 heights.
- Utilize this computed mean as our estimate for the population mean.

However, it's important to recognize that our sample of 100 trees is merely one of numerous potential samples, all of the same size, that could have been randomly chosen. Imagine if we took a series of distinct random samples from the same population, each with the same sample size:

Sample 1: Compute sample mean $\bar{x}$.

Sample 2: Compute sample mean $\bar{x}$.

Sample 3: Compute sample mean $\bar{x}$.

And so on...

With each new sample, we might obtain a different result because we are using a distinct subset of data to calculate the sample mean. This demonstrates that the sample mean is, in fact, a random variable.

The sample mean $(\bar{x})$ constitutes a random variable that possesses its own probability distribution known as the sampling distribution of the sample mean. This distribution of the sample mean will have a mean identical to μ and a standard deviation equivalent to the standard error $\left(\frac{\sigma}{\sqrt{n}}\right)$.

The standard error $\left(\frac{\sigma}{\sqrt{n}}\right)$ serves as the standard deviation of all conceivable sample means. In practical terms, we would typically take only one sample, but it is crucial to comprehend and quantify the variability that occurs between different samples.

It's important to note that s^2 represents the sample variance, while s stands for the sample standard deviation.

$$S_{\bar{x}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

Note: s^2 is the sample variance and s is the sample standard deviation.

Problem 1.8 Describe the distribution of the sample mean.

A population of fish has weights that are normally distributed with $\mu = 8$ lb. and $s = 2.6$ lb. If you take a sample of size $n = 6$, the sample mean will have a normal distribution with a mean of 8 and a standard deviation (standard error) of $\frac{26}{\sqrt{6}} = 1.061$ lb.

If you increase the sample size to 10, the sample mean will be normally distributed with a mean of 8 lb. and a standard deviation (standard error) of $\frac{26}{\sqrt{10}} = 0.822$ lb.

Notice, how the standard error decreases as the sample size increases.

The Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) states that as the sample size increases, the distribution of sample means approaches a normal distribution. In situations where the underlying distribution of the randomn variable is unknown or poorly understood, the CLT provides assurance that the distribution of sample means $(\bar{x})$ becomes increasingly normal with larger sample sizes (n).

As a rule of thumb, a commonly accepted guideline suggests that a sample size of $n \geq 30$ is typically sufficient for this approximation.

The Central Limit Theorem states that as the sample size grows, the sampling distribution of the sample mean becomes approximately normal, irrespective of the shape of the original population distribution.

Coefficient of Variation

Comparing standard deviations across diverse populations or samples proves challenging due to the influence of units of measurement. The coefficient of variation addresses this issue by representing the standard deviation as a percentage of the sample or population mean, making it a unitless measure.

Population data	Sample data
$CV = \frac{\sigma}{\mu} \times 100$	$CV = \frac{s}{\bar{x}} \times 100$

Problem 1.9 Fisheries biologists conducted research on Pacific salmon, focussing on their length and weight. They gathered a random sample and calculated the mean and standard deviation for both length and weight (provided below). Although the standard deviations are somewhat alike, the challenge lies in comparing variability due to the different units of measurement between lengths and weights. By computing the coefficient of variation for each variable, the biologists can effectively assess which variable exhibits the higher standard deviation, regardless of unit discrepancies.

	Sample mean	**Sample standard deviation**
Length	63 cm	19.97 cm
Weight	37.6 kg	19.39 kg

$$CV_L = \frac{19.97}{63} \times 100 = 31.7\% \qquad CV_W = \frac{19.39}{37.6} \times 100 = 51.6\%$$

There is greater variability in the weight of Pacific salmon as compared to its length.

Variability

Variability can be assessed through diverse approaches. The standard deviation measures the dispersion between individual data points in a sample, revealing the diversity among the sampled units. In contrast, the coefficient of variation evaluates point-to-point variability in a relative fashion, considering the mean and remaining independent of measurement units. Meanwhile, the standard error quantifies variability between different samples, portrayinhg the diversity across various samples within the sampling procedure. Typically, when dealing with a single sample, the standard error provides a means to quantify the inherent uncertainty in our sampling process.

Problem 1.10 Consider the following tally from 11 sample plots on Heiburg Forest, where X_i is the number of downed logs per acre. Compute the basic statistics for the sample plots.

Table 1.1 Sample data on number of downed logs per acre from Heiburg Forest

ID	X_i	X_i^2	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	Order
1	25	625	– 7.27	52.88529	4
2	35	1225	2.73	7.4529	6
3	55	3025	22.73	516.6529	10
4	15	225	– 17.25	298.2529	2
5	40	1600	7.73	59.7529	8
6	25	25	– 7.27	52.8529	5
7	55	3025	22.73	516.6529	11
8	35	1225	2.73	7.4529	7
9	45	2025	12.73	162.0529	9
10	5	25	– 27.27	743.6529	1
11	20	400	– 12.27	150.1819	3

(Contd...)

Sum	**355**	**14025**	**0.0**	**2568.1519**	
	$\sum_{i=1}^{n} X_i$	$\sum_{i=1}^{n} X_i^2$	$\sum_{i=1}^{n}(X_i - \bar{X})$	$\sum_{i=1}^{n}(X_i - \bar{X})^2$	

1. Sample mean:

$$\bar{x} = \sum \frac{x_i}{n} = \frac{355}{11} = 32.27$$

2. Median = 35
3. Variance:

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{14025 - \frac{(355)^2}{11}}{11-1}$$

$$= 256.82$$

4. Standard deviation:

$$s = \sqrt{s^2} = \sqrt{256.82} = 16.0256$$

5. Range: $55 - 5 = 50$
6. Coefficient of variation:

$$CV = \frac{s}{\bar{x}} \times 100 = \frac{16.0256}{32.27} \times 100 = 49.66$$

7. Standard error of the mean:

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{256.82}{11}} = 4.8319$$

1.4 GRAPHICAL REPRESENTATION

Data organization and summarization can be accomplished through both graphical and numerical methods. Tables and graphs offer a rapid overview of the collected information to facilitate the presentation of data utilized in the project. Although numerous types of graphics are available, this chapter will concentrate on a selected few commonly employed tools.

1.4.1 Pie Charts

Pie charts serve as effective visual aid for as quick understanding of relationships between categories. It's crucial to provide distinct labels for each category, and often, including the frequency or relative frequency can enhance understanding. Nonetheless, it's essential to exercise caution while incorporating numerous categories, as this can lead to confusion. Avoid overloading a pie chart with excessive information. The first pie chart offers a clear depiction of fish types in relation to the entire sample. Conversely, the second pie chart, with its abundance of categories, proves more challenging to interpret. Selecting the most appropriate graphical representation is imperative when conveying information to the reader.

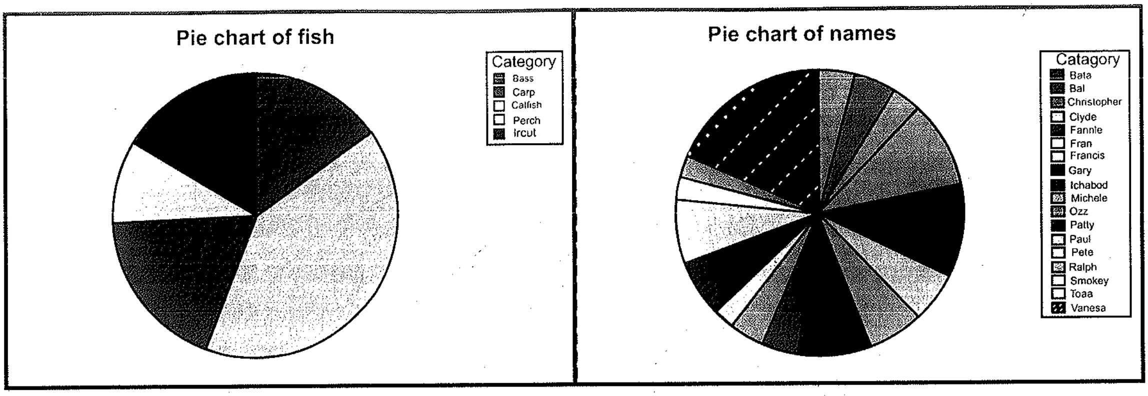


Fig. 1.5 Comparison of pie charts

1.4.2 Bar Charts and Histograms

Bar charts provide a graphical representation of the distribution of a qualitative variable, such as fish type, while histograms are used to depict the distribution of a quantitative variable, whether it's discrete or continuous, like bear weight.

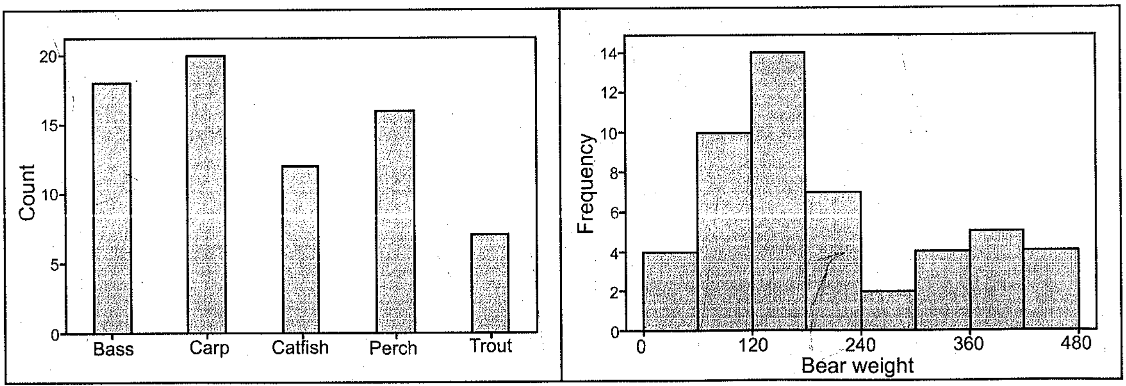


Fig. 1.6 Comparison of a bar chart for qualitative data and a histogram for quantitative data

In both the scenarios, the bars have uniform widths, and the y-axis is distinctly defined. For qualitative data, each category corresponds to a distinct bar. When working with continuous data, it's essential to establish consistent class widths by defining lower and upper class limits. There should be no gap between these classes, ensuring that each observation falls into precisely one class without any overlap.

1.4.3 Boxplots

A box plot, also known as a box-and-whisker plot, is a graphical representation that displays the distribution of a dataset's values. Boxplots employ the 5-number summary, which includes the minimum and maximum values along with the three quartiles, to depict the central tendency, spread, and data distribution. When combined with histograms, they provide a comprehensive description of the data, presenting both numerical and visual insights.

In cases where the data is symmetric, the distribution exhibits a bell-shaped and reasonably symmetrical pattern.

It consists of a rectangular box that represents the Interquartile Range (IQR), which spans the middle 50% of the data. Inside the box, a line or 'whisker' typically indicates the median (the middle value when the data is sorted). Two additional lines or whiskers extend from the box to represent the range of the data within a certain range from the median. These whiskers can indicate potential outliers or extreme values.

Box plots are useful for visualizing the central tendency, spread, and skewness of a dataset and can help identify outliers or data points that deviate significantly from the norm. They are a valuable tool in exploratory data analysis and are often used in statistics to compare and summarize data from different groups or populations.

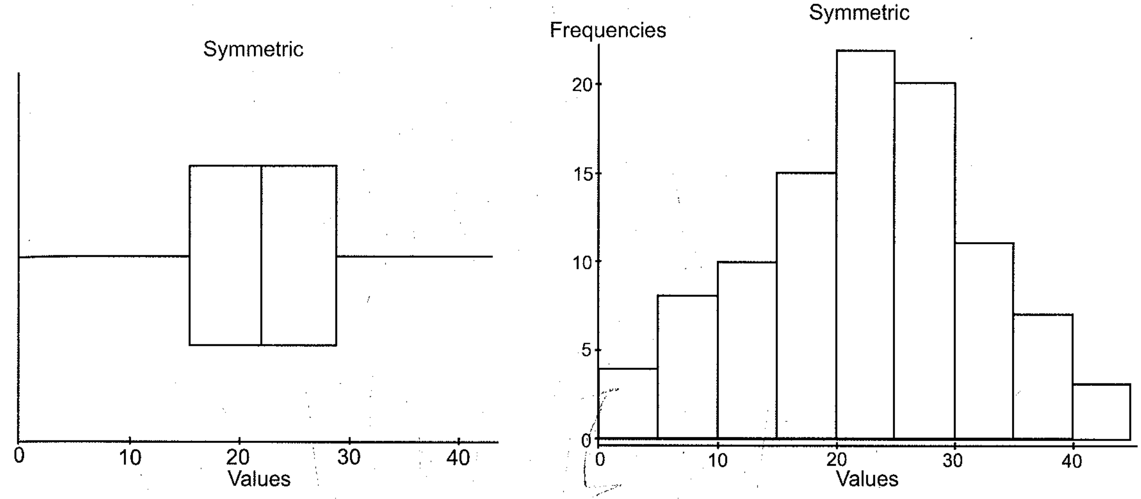


Fig. 1.7 A histogram and boxplot of a normal distribution

In the context of left-skewed distributions, the histogram visually appears to be shifted or 'pulled' towards the left side. When observing the corresponding boxplot, one can notice that the first quartile (Q1) is positioned farther from the median, as are the minimum values. Additionally, the left whisker of the boxplot is notably longer than the right whisker.

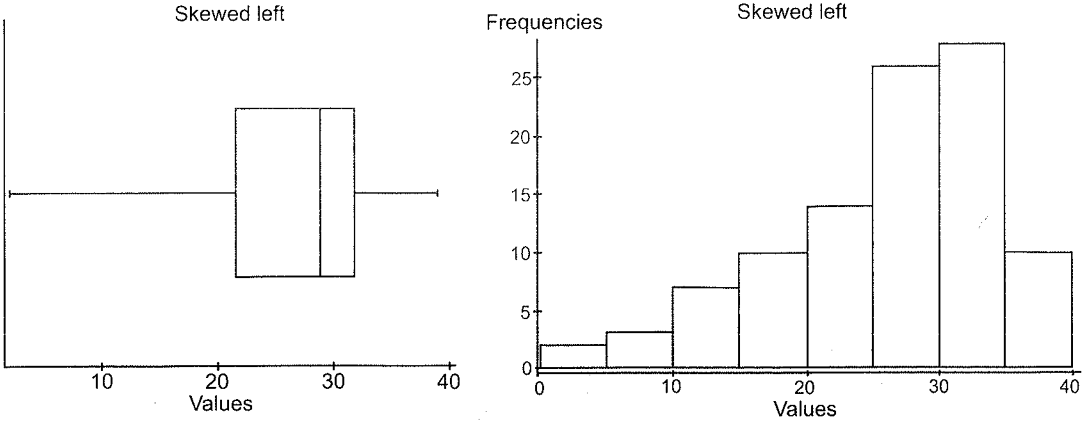


Fig. 1.8 A histogram and boxplot of a skewed left distribution

In the case of right-skewed distributions, the histogram exhibits a visual shift or 'pull' towards the right side. When examining the corresponding boxplot, you will notice that the third quartile (Q3) is positioned farther from the median, as is the maximum value. Furthermore, the right whisker in the boxplot is noticeably longer than the left whisker.

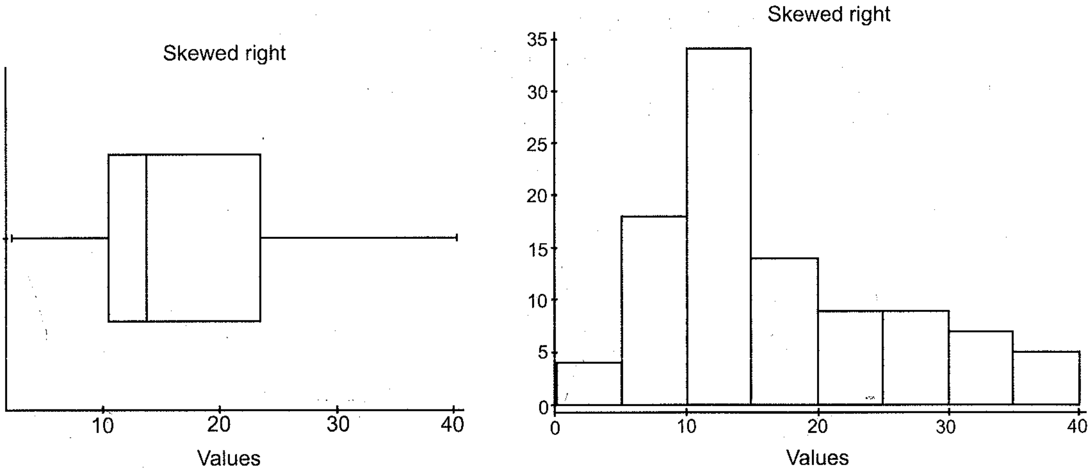


Fig. 1.9 A histogram and boxplot of a skewed right distribution

PROBLEMS

1. A sample contains the following data values: 1.50, 1.50, 10.50, 3.40, 10.50, 11.50, and 2.00. Find the mean, median, and variance of this data?
2. **Set 1:** 5, 13, -2, -1, 19, 27

   Set 2: 6, – 3, 23, 15, *m*, 1

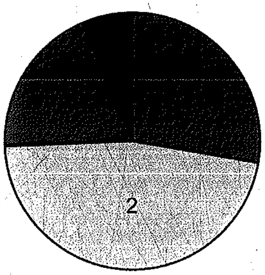
   Find out should the value of *m* if the ranges of both sets of numbers are equal?
3. Find the coefficient of variation of the following sample set of numbers.

   {1, 5, 6, 8, 10, 40, 65, 88}

4. Brian pouts a question 60 people about their favourite colours and separates their answers into 5 categories. The results are shown in the table below. Draw a pie chart to display the results.

Colour	Red	Blue	Green	Yellow	Other
Frequency	10	13	24	5	8

5. A survey was done for 90 people asking them. How many bathrooms are there in their homes? Freddie makes a pie chart to display the results of this survey.



(a) The number of bathrooms in a house was most common question in this survey?

(b) Find out the number of people who have 1 bathroom in their house.

(c) If Freddie picks someone at random from the group surveyed, what will be the probability that he chooses 1 person with exactly 1 bathroom in their home?

6. Construct a box plot for the following dataset.

3, 5, 8, 8, 9, 11, 12, 12, 13, 13, 163, 5, 8, 8, 9, 11, 12, 12, 13, 13, 16

CHAPTER

2

PROBABILITY THEORY AND PROBABILITY DISTRIBUTION

Chapter Outline

• Introduction • Types of Random Variables • Jointly Distributed Random Variables • Independent and Dependent Random Variables • Expectation of the Sum of the Random Variables • Expectation of the Product of the Random Variables • Probability Distribution

2.1 INTRODUCTION

In random experiments, we are interested in the numerical outcomes i.e., numbers associated with the outcomes of the experiment. For example, when 50 coins are tossed, we ask for the number of heads. Whenever we associate a real number with each outcome of trial, we are dealing with a function whose range is the set of real numbers we ask for such a function is called a random variable (*r.v.*) chance variable, stochastic variable or simply a variable.

Definition: Quantities which vary with some probabilities are called random variables.

Definition: By a random variable we mean a real number associated with the outcomes of arandom experiment.

Example 2.1 Suppose two coins are tossed simultaneously then the sample space is: $S = \{\text{HH, HT, TH, TT}\}$. Let X denote the number of heads, then if $X = 0$ then the outcome is {TT} and $P(X = 0) = \frac{1}{4}$. If X takes the value 1, then outcome is {HT, TH} and $P(X = 1) = \frac{2}{4}$. Next if X takes the value 2 then the outcome is {HH} and $P(X = 2) = \frac{1}{4}$. The probability distribution of this random variable X is given by the following table:

$X = x$	0	1	2	**Total**
$P(X = x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$	1

Example 2.2 Out of 24 mangoes 6 are rotten, 2 mangoes are drawn. Obtain the probability distribution of the number of rotten mangoes that can be drawn:

Let X denote the number of rotten mangoes drawn then X can take values 0, 1, 2.

$$P(X = 0) = \frac{{}^{18}C_2}{{}^{24}C_2} = \frac{18 \times 17}{24 \times 23} = \frac{51}{92};$$

$$P(X = 1) = \frac{{}^{18}C_1 \times {}^{6}C_1}{{}^{24}C_2} = \frac{18 \times 6}{\dfrac{24 \times 23}{1 \times 2}} = \frac{9}{23}$$

and

$$P(X = 2) = \frac{{}^{6}C_2}{{}^{24}C_2} = \frac{6 \times 5}{24 \times 23} = \frac{5}{92}$$

X = x	0	1	2	**Total**
***P* (X = x)**	$\frac{51}{92}$	$\frac{9}{23}$	$\frac{5}{92}$	1

2.2 TYPES OF RANDOM VARIABLES

There are two types of random variables:

(i) Discrete random variables
(ii) Continuous random variables

Distribution function: Let X be a one-dimensional random variable. The function F be defined for all x, by the equation $F_X(x) = P(X \le x)$ is called the cumulative distribution function of X.

Note:

1. We write *c.d.f.* For cumulative distribution function. Only *d.f.* is written instead of *c.d.f.*
2. Suffix X in F_X is used to emphasize the fact that the distribution function is associated with the value at X. when the underlying variate is clear from the context, we shall simply write $F(x)$ instead of $F_X(x)$.
3. Tail events let 'x' be any real number then the events $|X < x|$ and $|X > x|$. $|X \ge x|$ are called tail events. For distinction, we may label them open, closed, upper and lower tails. Often, simple *r.v.*'s are expanded as linear combination of tail events.

Some Properties of a. *c.d.f.*:

(1) $P\{a < X \le b\} = F(b) - F(a)$	Interval property
(2) $0 \le F(x) \le 1, \forall \times \in \mathbb{R}$	Boundedness property

(3) F is non-decreasing i.e., if $x \leq y$, then $F(x) \leq F(y)$ Monotone increasing property

(4) $F(-\infty) = 0, F(+\infty) = 1$ Limits property

(5) F is continuous from the right each point (*a*) i.e., $F(a+) = F(a)$ Right continuous property

(6) $F(a+) - F(-a) = P(X = a)$ Jump discontinuity

Conditions (2),(3), (4) and (5) are necessary as well as sufficient for F to be c.d.f. on R.

Example 2.3 Give reasons why the following graphs of F given below does not represent a distribution function.

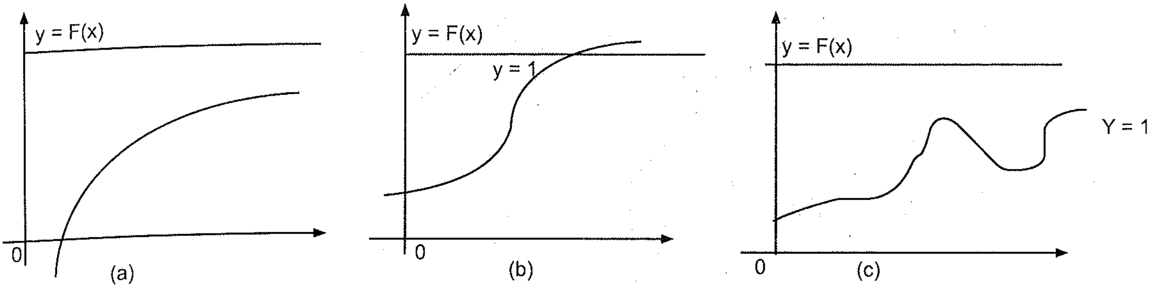


Fig. 2.1

Solution:

(a) $F(x) < 0$ – ve for some x

(b) $F(x) > 1$ for some x

(c) F is not non-decreasing i.e., sometimes F is decreasing also

2.2.1 Discrete Random Variables

Quantities which can take only integer values are called discrete random variables.

Examples

The number of children in a family of a colony.

The number of rooms in the houses of a township.

Probability mass function: Probability distribution

Definition: Let X be a discrete random variable taking value x, $x = 0, 1, 2, 3, \ldots$ then $P(X = x)$ is called the probability mass function of X and it satisfies the following

(*i*) $P(X = x) \geq 0$

(*ii*) $\sum_{x=0}^{\infty} P(X = x)$

Discrete distribution function

A *r.v.* X is said to be discrete, if there exist a countable number of points $x_1, x_2, x_3, \ldots$ and. $p(xi) \geq 0$ such that

$$\sum_{x=1}^{\infty} p(x_i) = 1 \; F(x) = \sum_{x_i \leq x} p(x_i)$$

Finite equiprobable space (Uniform space)

A finite equiprobable space is finite probability distribution where each sample point $x_1, x_2, x_3, \ldots x_n$ has the same probability for all *i i.e.*,

$$P(X = x_i) = p_i = \text{a constant for all } i \text{ and}$$

$$\sum_{x_i \leq x} p_i = 1$$

Example 2.4 A random variable X has the following probability distribution:

x	0	1	2	3	4	5	6	7	8
P(x)	k	$3k$	$5k$	$7k$	$9k$	$11k$	$13k$	$15k$	$17k$

(*a*) Determine the value of k

(*b*) Find $P(X < 4)$, $P(X \geq 5)$, $P(0 < X < 4)$

(*c*) Find the *c.d.f.*

(*d*) Find the largest value of x for which $P(X \leq x) < 0.5$.

Solution.

(*a*)
$$\sum_{x=1}^{\infty} p(x_i) = 1$$

$$k + 3k + 5k + 7k + 9k + 13k + 15k + 17k = 1$$

$$k = \frac{1}{81}$$

(*b*)
$$P(X < 4) = \frac{16}{81}$$

$$P(X \geq 5) = \frac{56}{81}$$

$$P(0 < X < 4) = \frac{15}{81}$$

(*c*)
$$\begin{cases} 0 & x < 0 \\ \dfrac{1}{81} & 0 \le x < 1 \\ \dfrac{9}{81} & 2 \le x < 2 \\ \dfrac{9}{81} & 2 \le x < 3 \\ \dfrac{16}{81} & 3 \le x < 4 \\ \dfrac{25}{81} & 4 \le x < 5 \\ \dfrac{36}{81} & 5 \pounds x < 6 \\ \dfrac{49}{81} & 6 \le x < 7 \\ \dfrac{64}{81} & 7 \le x < 8 \\ 1 & x \ge 8 \end{cases}$$

(*d*)
$$F(5) = 0.44$$
$$F(6) = 0.61$$

Therefore, the largest value of $x = 5$ for $P(X \le x) < 0.5$.

Expectation: The behaviour of *r.v.* either discrete or continuous is completely characterized by the distribution function $F(x)$ or density $f(x)[P(x_i)$ in discrete case]. Instead of a function, a more compact description can be made by a single number such as mean (expectation), median and mode known as measures of central tendency of the *r.v.* X.

Expectation or median or expected value of a *r.v.* X denoted by $E[X]$ or μ, is defined as

$$E[X] = \left\{ \begin{array}{l} \displaystyle\sum_{-\infty}^{\infty} x\, P(X = x), \text{ if } X \text{ is discrete} \\ \displaystyle\int_{-\infty}^{\infty} x f(x)\, dx, \text{ if } X \text{ is continuous} \end{array} \right\}$$

Definition

Variance: Variance characterizes the variablility in the distributions, since two distributions with same mean can still have different dispersion of data about their means,

Variance of *r.v.* X is for X discrete

$$\sigma^2 \equiv E[(X-\mu)^2] \equiv \Sigma(X-\mu)^2 f(x) \text{ for } X \text{ discrete}$$

$$\sigma^2 \equiv E[(X-\mu)^2] \equiv \int_{-\infty}^{\infty} (X-\mu)^2 f(x)\, dx \text{ for } X \text{ is continuous}$$

Standard Deviation: Standard deviation denoted by σ (*S.D.*) is the positive square root of variance.

$$\begin{aligned}\sigma^2 = E[(X-\mu)^2] &= \Sigma(X-\mu)^2 f(x) = \Sigma_x (x^2 - 2\,\mu x + \mu^2)\, f(x)\\ &= \Sigma_x x^2 f(x) - 2\mu\, \Sigma_x xf(x) + \mu^2\, \Sigma_x f(x)\\ &= E(X^2) - 2\mu\cdot\mu + \mu^2\cdot 1 = E(X^2) - \mu^2\end{aligned}$$

Since, $\mu = \Sigma_x xf(x),\ \Sigma_x f(x) = 1$

2.2.2 Continuous Random Variables

Let X be a continuous random variable taking value x, $a \le x \le b$ then $f(x) = P(X = x)$ is defined as the *p.d.f.* of X and satisfies the following:

(i) $f(x) \ge 0$

(ii) $\int_a^b f(x)\,dx = 1$

Note:

1. For a continuous variate, point probabilities are zero.
2. Area under the probability curve $y = f(x)$ is unity; the fact $f(x) \ge 0$ implies the graph $f(x)$ is above x = axis.
3. Area under the probability curve $y = f(x)$ bounded by $x = a$, $x = b$ is simply $P(a \le x \le b)$.
4. Relation between *p.d.f.* and *c.d.f.*: The density f and *c.d.f.* F are always connected by (a) $F(x) = \int_{-\infty}^{x} f(x)\,dx\ \forall\ x \in R$ (b) $\frac{d}{dx}[F(x)] = f(x)\ \forall\ x \in R$.

Moments: If the range of the probability density function is from $-\infty$ to ∞, the r^{th} moment about origin is defined as $\mu_r' = \int_{-\infty}^{\infty} x^r f(x)\,dx$.

The r^{th} moment about any arbitrary origin 'a' is $\mu_r' = \int_{-\infty}^{\infty} (x-a)^r f(x)\,dx$

The mean is given by (taking moment about $x = 0$) $\mu_1' = \int_{-\infty}^{\infty} xf(x)\,dx$

The variance σ^2 is given by $\sigma^2 = \mu_2' - \mu_1'^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \left[\int_{-\infty}^{\infty} xf(x)dx\right]^2$

2.3 JOINTLY DISTRIBUTED RANDOM VARIABLES

When the outcome of a random experiment can be characterized in more than one way, the probability density is a function of more than one variate.

Example 2.5 When a card is drawn from an ordinary deck, it may be characterized according to its suit in some order viz., say clubs, diamonds, hearts and spades and Y be a variate that assumes the values 1, 2, 3, ..., 13 which correspond to the denominations: Ace, 2, 3, ..., 10, J, Q, K. Then (X, Y) is a 2-dimensional variate. The probability of drawing a particular card will be denoted by $f(x, y)$ and if each card is equi-probable of being drawn, the density of (X, Y) is

$$f(x, y) = \frac{1}{52} \ \forall\ 1 \le x \le 4\ \forall\ 1 \le y \le 13$$

Trails whose outcomes can be characterized by two (three) variates give rise to bivariate (tri-variate) distributions, etc. Extensions to n-variate distributions are fairly straight forward.

2.3.1 Joint Distribution Function and its Properties

Let (X, Y) be a random vectoror random variable on the probability space. The joint *c.d.f.* of X and Y is denoted by $F_{X,Y}$ and is defined by $F_{X,Y}(x, y) = P(X \le x, Y \le y), x, y \in R$.

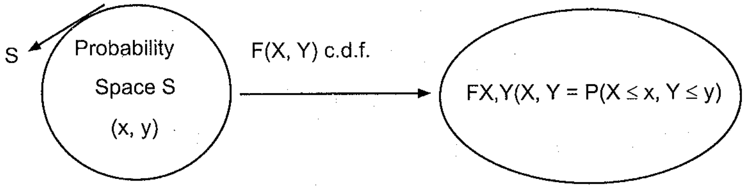


Fig. 2.2 Joint Distribution Function

A joint *c.d.f.* of two variates has the following properties:

1. **Non-negativity and Boundedness:**

$$0 \le F_{X,Y}(x, y) \le 1, \text{ for every } x, y \in \mathbb{R}.$$

2. **Monotonicity:** The *c.d.f.* 'F' is monotonically non-decreasing function in each of the individual variables, *i.e.*,

   (i) $F(a, y_2) \ge F(a, y_1)$, if $y_2 \ge y_1$

   (ii) $F(x_2, b) \ge F(x_1, b)$, if $x_2 \ge x_1$.

3. **Rectangle rule:** Let a, b, c, d be any real numbers with $a < b$ and $c < d$.

   Then, $P(a < X \le b, c < Y \le d) = F(b, d) + F(a, c) - F(b, c) - F(a, d)$

4. **Individual limits:**

   (i) $\lim\limits_{x \to -\infty} = F(-\infty, y) = 0$ (ii) $\lim\limits_{y \to -\infty} F(x, y) = F(x, -\infty) = 0$

5. **Double limit:** $\lim\limits_{x \to \infty, y \to \infty} F(x, y) = 1$ *i.e.*, $F(\infty, \infty) = 1$ formally.

   Note: We don't claim $F(\infty, y) = 1$ or $F(x, \infty) = 1$.

6. **Individual Continuity:** F is continuous from the right in each of its individual variables. *i.e.*,

   (i) $\lim_{x \to a^+} F(x, y) = F(a, y)$

   (ii) $\lim_{y \to b^+} F(x, y) = F(x, b)$

7. If the density function $f(x, y)$ is continuous at (x, y), then $\frac{\partial^2 F}{\partial x \partial y} = f(x, y)$.

2.3.2 Joint Discrete Distribution Function

The joint *c.d.f.* of X and Y is said to be discrete if there exists a non-negative function P such that P vanishes everywhere except a finite or countably infinite number of points in the plane and at such points (x, y) so that $P(x, y) = P(X = x, Y = y)$, for all $x, y \in R$.

Let X and Y have a joint discrete distribution. A function P which does not vanish on the set $\{(x_i, y_j)$ such that $i, j = 1, 2, 3, \ldots\}$ and satisfies the following properties:

(i) $P(x_i, y_j) \geq 0$ for all $i, j = 1, 2, 3, \ldots\ldots$ and

(ii) $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} P\left(x_i, y_j\right) = 1$ is called joint probability (mass) function of X and Y or simply the joint probability function.

Individual and Marginal Probability Functions

Let X and Y be two jointly distributed variables with joint discrete density $P(x, y)$, the individual variates X and Y themselves are random variables.

The individual distributions of X and Y are called marginal distributions of X and Y:

(i) The Marginal probability function for X is denoted by $P_X(x)$ or $P(x)$ and is given by

$$P(x) = P(X = x) = \Sigma_y P(X = x, Y = y) = \Sigma_y P(x, y)$$

(ii) The marginal probability function for Y is denoted by $P_Y(y)$ and is given by

$$P(y) = P(Y = y) = \Sigma_x P(X = x, Y = y) = \Sigma_x P(x, y)$$

Note: It is convenient to display the probability function of a bivariate distribution in a rectangular array, in which the row totals and column totals provide the marginal probability functions of X and Y, respectively.

Table 2.1 Conditional Probability Table

y x	y_1	y_2	y_3	...	y_j	...	y_m	...	$P(X = xi)$
x_1	P_{11}	P_{12}	P_{13}	...	P_{1j}	...	P_{1m}	...	$P(x_1)$
x_2	P_{21}	P_{22}	P_{23}	...	P_{21}	...	P_{2m}	...	$P(x_2)$
x_3	...	...	...	...	...	...	...	...	...
.	...	...	...	...	...	...	...	...	...
x_i	P_{i1}	P_{i2}	P_{i3}	...	P_{ij}	...	P_{im}	...	$P(x_i)$
.	.	.	.	.	.	.	.	.	.
x_n	P_{n1}	P_{n2}	P_{n3}	...	P_{nj}	...	P_{nm}	...	$P(x_n)$
.	...	...	...	...	...	...	...	...	...
$P(Y = y_i)$	$P(y_1)$	$P(y_2)$	$P(y_3)$	...	$P(y_j)$	...	$P(y_m)$	...	1

We have here, $P_{ij} = P(X = x_i, Y = y_j)$;

$$P(x_i) = \Sigma_j P_{ij} \quad \text{and} \quad P(y_j) = \Sigma_i P_{ij}$$

Conditional Probability Functions (cond. P. f.)

Let X and Y have joint discrete distribution with associated probability function P. Let the possible values of X be $\{x_1, x_2, x_3, ..., x_i, ...\}$ and those of Y be $\{y_1, y_2, y_3, ..., y_j, ...\}$ respectively.

The conditional probability function of X, given $Y = y_j$ denoted by $P_{X/y_j}\left({x_i}/{y_j}\right)$ is defined by

$$P_{X/y_j}\left({x_i}/{y_j}\right) = {P(x_i, y_j)}/{P_Y(y_j)} \quad \text{for } i, j = 1, 2, 3, ...$$

$$= 0 \text{ if } P_Y(y_j) = 0$$

The conditional probability function of Y, given $X = x_i$ denoted by $P_{Y/x_i}\left({y_j}/{x_i}\right)$ is defined by

$$P_{Y/x_i}\left({y_j}/{x_i}\right) = {P(x_i, y_j)}/{P_X(x_i)} \quad \text{for } i, j = 1, 2, 3, ...$$

$$= 0 \text{ if } P_X(x_i) = 0$$

Therefore, $P(x_i, y_j) = P(X = x_i, Y = y_j)$;

$$P(Y = y_j) = P_Y(y_j) \quad \text{and} \quad P(X = x_i) = P_X(x_i)$$

2.3.3 Joint Continuous Distribution Function

A 2-dimensional random vector (X, Y) is called a continuous random vector if there exists a function $f(x, y) \geq 0$ such that for $-\infty < x, y < \infty$, the *c.d.f.* of (X, Y) given by $F(x, y) = \int_{-\infty}^{x}\left[\int_{-\infty}^{y} f(u, v)\, dv\right] du$ is continuous. The function $f(x, y)$ is called the joint *p.d.f* of (X, Y).

Some properties of joint density: Let $f(x, y) \geq 0$ be the joint *p.d.f* of continuous random vector (X, Y) and $F(x, y)$ be the *c.d.f.* of (X, Y) then it holds the following properties:

(i) $\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} f(x, y)\, dx\right] dy = 1$

(ii) $P\{a < X \leq b, c < Y \leq d\} = \int_{a}^{b}\int_{c}^{d} f(x, y)\, dy dx$

(iii) $f(x, y) = \dfrac{\partial^2 F(x, y)}{\partial x \partial y}$

Individual or Marginal Distributions: Let (X, Y) be a continuous random vector with joint *c.d.f.* '*F*' and joint *p.d.f.* '*f*'. Then $F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u, v)\, dv\, du$.

Definition: Let (X, Y) be a 2-dimesnional continuous random vector with joint *p.d.f.* $f(x, y)$. Then the individual or marginal distribution of X and Y are defined by the *p.d.f.*'s

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

On observation, we have $P(a \leq X \leq b) = \int_{a}^{b} f_X(x) dx = \int_{a}^{b}\left[\int_{-\infty}^{\infty} f(x, y)\, dy\right] dx$.

Conditional Distribution Function: The conditional *c.d.f.* of a variate X, given $Y = y$, written $F_{X/Y}(X/Y) = \lim_{\epsilon \to 0^+} P\{X \leq x/y - \epsilon \leq Y \leq y + \epsilon\}$.

The conditional *p.d.f.* of X given $Y = y$, written $f_{X/Y}(x/y)\ \forall\ x \in \mathbb{R}$ is a non-negative function satisfying $F_{X/Y}(X/Y) = \int_{-\infty} f_{X/Y}\left(t/y\right) dt\ \ \forall\ x \in \mathbb{R}$.

Note: The conditional *p.d.f.* $f(x/y)$ is given by $f_{X/Y}(x/y) = f(x,y)/f_Y(y)$ where $f_Y(y)$ is the marginal *p.d.f.* of Y, $f_Y(y) > 0$, and is continuous.

Let X be a discrete random variable taking value x, $x = 0, 1, 2, 3, \ldots$ then the mathematical expectation of X is denoted by $E(X)$ [read it as expected value of X] and is defined as

$$E(X) = \sum_{x} xP(X = x)$$

Similarly, if k is any positive integer then, $E(X^k) = \Sigma_x x^k P(X = x)$.

Similarly, if X is a continuous random variable taking the value x, $-\infty < x < \infty$ with $f(x)$ as the probability density function then $E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$ and $E(X^k) = \int_{-\infty}^{\infty} x^k f(x)\,dx$.

Definition: $E(X)$ is also called mean or arithmetic mean of X denoted by μ.

Definition: If X is a random variable, then the variance of X is denoted by $V(X)$ and is defined as

$$V(X) = E[(X - E(X))^2]$$

This can be simplified as $V(X) = E(X^2) - [E(X)]^2$.

Notation: The variance is denoted by $\sigma^2 = V(X)$.

Standard deviation: The positive square root of variance is defined as standard deviation and is denoted by σ. Therefore, $\sigma = \sqrt{V(X)}$.

2.4 INDEPENDENT AND DEPENDENT RANDOM VARIABLES

Two random variables X and Y are called independent if for every pair of real number 'x' and 'y', the two events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent. That is we can express as:

$$P\{X \leq x, Y \leq y\} = P\{X \leq x\}\, P\{Y \leq y\}$$

In terms of distribution functions:

$$F(x, y) = F_X(x)\, F_Y(y)$$

and also $f(x, y) = f_X(x)\, f_Y(y)$ [if densities exist]

Variates which are not independent are called dependent variates or dependent random variables.

2.5 EXPECTATION OF THE SUM OF THE RANDOM VARIABLES

The mathematical expectation of a sum of a number of random variables is equal to the sum of their expectations.

Proof: Let us consider the random variables x and y. Let x assume the values x_i for all $i = 1, 2, 3, \ldots, m$ and y the values y_j for all $j = 1, 2, 3, \ldots, n$ with respective probabilities P_i and P_j. The sum $x + y$ is a random variable which can take $m \times n$ values,

$x_i + y_j$ for $i = 1, 2, 3, ..., m$ and $J = 1, 2, 3, ..., n$ with probabilities P_{ij}.

Hence its Expectation is

$$
\begin{aligned}
E(x+y) &= \sum_{j=1}^{n}\sum_{i=1}^{m}(x_i + y_j)\,P_{ij} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{m} x_i\,P_{ij} + \sum_{j=1}^{n}\sum_{i=1}^{m} y_i\,P_{ij} \\
&= \sum_{i=1}^{m} x_i\left(\sum_{j=1}^{n} P_{ij}\right) + \sum_{j=1}^{n} y_j\left(\sum_{i=1}^{m} P_{ij}\right) \\
&= \Sigma_i\, x_i\, P_i + \Sigma_j\, y_j\, P_j = E(x) + E(y)
\end{aligned}
$$

$$\therefore \qquad E(x+y) = E(x) + E(y)$$

Since, $\displaystyle\sum_{j=1}^{n} P_{ij} = P_i$ and $\displaystyle\sum_{i=1}^{m} P_{ij} = P_j$.

By generalization of the above theorem, we have

$E(x_1 + x_2 + x_3 + ... + x_n) = E(x_1) + E(x_2) + E(x_3) + ... + E(x_n)$

This completes the proof of the theorem.

2.6 EXPECTATION OF THE PRODUCT OF THE RANDOM VARIABLES

The mathematical expectation of the product of a number of independent random variates is equal to product of their expectations.

Proof: Let us use the notation as $E(xy) = \Sigma_j\, \Sigma_i\, x_i\, y_j\, P_{ij}$

Since the variates are independent, by the law of compound probabilities we have

$$P_{ij} = P_i P_j$$

$$\sum_j\sum_i x_i P_i y_j P_j = \sum_i x_i P_j \sum_j y_i P_j = \sum_i p_i x_i E(y) = E(y)\sum_i p_i x_i = E(x)E(y)$$

The theorem can be generalized for a number of independent random variates such that

$$E(x_1 \,.\, x_2 \,.\, x_3 \,. \dots .\, x_n) = E(x_1) \,.\, E(x_2) \,.\, E(x_3) \,. \dots .\, E(x_n)$$

This completes proof of the theorem.

Note: $E(x, y) = E(x)\,E(y)$ does not guarantee the independent of x and y.

One can easily verify that the following expectations:

(a) $E(a) = a$

(b) $E(aX) = a\,E(X)$

(c) $E(aX \pm bY) = aE(X) \pm bE(Y)$

(d) $E(aX + b) = a\,E(X) + b$

(e) $V(a) = 0$

(f) $V(aX \pm b) = a^2\,V(X)$

(g) $V(x) = E(x^2) - [E(x)]^2$

(h) $V(aX + bY) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\sigma_{XY}$

Example 2.6 Two coins are tossed simultaneously. Let X denote the number of heads, Find $E(X)$ and $V(X)$?

Solution:

$X = x$	0	1	2	Total
$P(X = x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$	1

Mean: $$\mu = E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{2}{4} + 2 \cdot \frac{1}{4} = 1$$

Variance: $$\sigma^2 = V(X) = E(X^2) - [E(X)]^2$$

$$= 0^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{2}{4} + 2^2 \cdot \frac{1}{4} - (1)^2$$

$$= \frac{2}{4} + 1 - 1$$

$$= \frac{1}{2}$$

Hence, the solution.

Example 2.7 If it rains, a dealer in raincoats earns ₹ 500 per day and if it is fair, he loses ₹ 50 per day. If the probability of a rainy day is 0.4. Find his average daily income?

Solution:

$X = x$	500	− 50	Total
$P(X = x)$	0.4	0.6	1

Average, $$E(X) = 500\,(0.4) + (-50)\,(0.6)$$

$$= 200 - 30$$

$$= ₹\ 170/-$$

Hence, the solution.

2.7 PROBABILITY DISTRIBUTION

2.7.1 Binomial Distribution

This distribution was discovered by James Bernoulli. This is a discrete distribution. It occurs in cases of repeated trials such as students writing an examination, births in a hospital, etc. Here, all the trials are assumed to be independent and each trial has only two outcomes namely success and failure.

Let an experiment consist of 'n' independent trials. Let it succeed 'x' times. Let 'p' be the probability of success and 'q' be the probability of failure in each trial.

$$\therefore \qquad p + q = 1$$

The probability of getting x successes = $p \cdot p \cdot p \ldots\ldots p(x \text{ times}) = p^x$

The probability of getting $(n - x)$ failures = $q \cdot q \cdot q \ldots\ldots q(n - x \text{ times}) = q^{n-x}$

$\therefore$ From multiplication theorem, the probability of getting x successes and $(n - x)$ failures is $p^x q^{n-x}$.

This is the probability of getting x successes in one combination. There are such ${}^nC_x$ mutually exclusive combinations each with probability $p^x q^{n-x}$.

$\therefore$ From addition theorem the probability of getting x success in ${}^nC_x p^x q^{n-x}$.

Notation: $b(x; n, p)$ denotes a binomial distribution with x successes, n trials and with p as the probability of success.

$$\therefore \qquad b(x; n, p) = {}^nC_x\, p^x q^{n-x}, x = 0, 1, 2, 3, \ldots, n$$

Parameters of Binomial Distribution

In $b(x; n, p)$ there are 3 constants viz., n, p and q. Since, $q = 1 - p$, hence there are only 2 independent constants namely n and p. These are called the parameters of binomial distribution.

Note: Since $b(x; n, p)$ is same as the $(x + 1)^{th}$ term in the binomial expansion of $(q + p)^n$, hence this distribution is called the 'Binomial Distribution'.

Mean of the Binomial Distribution

$$\text{Mean} = \mu = \sum_{x=0}^{n} x \cdot b(x; n, p)$$

$$= \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!\,(n-x)!} p^x p^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!\,(n-x)!} p^{x-1} p^{n-x}$$

Put $y = x - 1$, $\therefore$ $x = 1 + y$

When $x = 1$ implies $y = 0$

$x = 1$ implies $y = x - 1$

$$= np \sum_{y=0}^{n-1} {}^{(n-1)}C_y \; p^y q^{n-y-1}$$

$$= np(q + p)$$

So, m = np is Mean of Binomial distribution.

Variance of the Binomial Distribution

$$E(X)^2 = \sum_{x=0}^{n} x^2 b(x,n,p)$$

$$= \sum_{x=0}^{n} [x(x-1)+x] b(x,n,p)$$

$$= \sum_{x=0}^{n} x(x-1) b(x,n,p) + \sum_{x=0}^{n} x b(x,n,p)$$

$$= \sum_{x=0}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x q^{(n-x)} + \mu$$

$$= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!} p^x q^{(n-x)} + np$$

$$= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} p^{(x-2)} q^{(n-x)} + np$$

Put $y = x - 2$ $\therefore$ $x = 2 + y$

When $x = 2$ implies $y = 0$

When $x = n$ implies $y = n - 2$

$$= n(n-1)p^2 \Sigma_{y=0}^{n-2} \frac{(n-2)!}{y!(n-2-y)!} p^y q^{(n-2-y)} + np$$

$$= n(n-1)p^2 \sum_{y=0}^{n-2} {}^{(n-2)}.C_y p^y q^{(n-2-y)} + np$$

$$= n(n-1)p^2(q+p)^{n-2} + np$$

$$E(X^2) = n(n-1)p^2 + np$$

$$\sigma^2 = E(X^2) - [E(X)]^2$$

$$= n(n-1)p^2 + np - n^2p^2$$

$$= np(1-p)$$

$$\sigma^2 = npq$$

∴ The variance of binomial distribution is npq. The standard deviation is $\sigma = \sqrt{npq}$.

Moments of Binomial Distribution

$$\text{Mean } \mu'_1 = \mu = E[X] = \Sigma_x xP(x)$$

$$= \sum_{x=0}^{n} x\, {}^nC_x p^x q^{(n-x)}$$

$$= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{(n-x)}$$

$$= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!\,(n-x)!} pp^{(x-1)} q^{(n-x)}$$

$$= np\sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!\,(n-x)!} p^{(x-1)} q^{(n-x)}$$

$$= np(p+q)^{(n-1)} = np(1)^{(n-1)} = np$$

$E[X^2] = \mu'_2 = \Sigma_x x^2P(x)$ by definition

$$= \sum_{x=0}^{n} x^2\, {}^nC_x p^x q^{(n-x)}$$

$$\mu'_2 = \sum_{x=0}^{n} \{x(x-1)+x\}\, {}^nC_x p^x q^{(n-x)}$$

$$\mu'_2 = \sum_{x=0}^{n} x(x-1)\, {}^nC_x p^x q^{(n-x)} + np$$

Consider $$\mu''_2 = \sum_{x=0}^{n} x(x-1)\, {}^nC_x p^x q^{(n-x)}$$

$$\mu''_2 = \sum_{x=0}^{n} x(x-1) \frac{n!}{x!\,(n-x)!} p^x q^{(n-x)}$$

$$\mu''_2 = \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!} p^x q^{(n-x)}$$

$$\mu''_2 = \sum_{x=2}^{n} \frac{n(n-1)\,(n-2)!}{(x-2)!(n-x)!} p^2 p^{(x-2)} q^{(n-x)}$$

$$\mu''_2 = n(n-1)\,p^2(q+p)^{(n-2)}$$

$$\mu''_2 = n(n-1)\,p^2$$

$$\therefore \qquad \mu'_2 = n(n-1)\,p^2 + np$$

$$\begin{aligned} \mu_2 \text{ (variance)} &= \mu'_2 - \mu'^2_1 \\ &= n(n-1)p^2 + np - (np)^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np(1-p) \\ &= npq \quad [\text{since, } 1-p=q] \end{aligned}$$

$$\therefore \qquad np > npq \quad [\text{since } q \text{ is a fraction}]$$

$$\text{Mean} > \text{Variance}$$

Similarly $\mu_3 = npq(1-2p)$

Hence, $\beta_1 = \dfrac{\mu_3^2}{\mu_2^2} = \dfrac{n^2p^2q^2(1-2p)^2}{(npq)^2} = \dfrac{(1-p-p)^2}{npq} = \dfrac{(q-p)^2}{npq}$

When $p = \dfrac{1}{2} = q$

Therefore, $\beta_1 = 0$

Case (i) when $p = \dfrac{1}{2}, \beta_1 = 0$

Case (ii) when $n \to \infty$ then $\beta_1 = 0$

$$\text{Standard deviation} = \sqrt{npq}$$

$$\text{Skewness} = \frac{1-2p}{\sqrt{npq}}$$

Moment Generating Function of Binomial Distribution

Let X be a variable following binomial distribution, then

$$M_x(t) = E(e^{tx}) = \sum\nolimits_{x=0}^{n} e^{tx} p(x)$$

$$= \sum_{x=0}^{n} e^{tx}\ {}^{n}C_{x} p^{x} q^{(n-x)}$$

$$= (q + pe^{t})^{n}$$

M.G.F about Mean of Binomial Distribution

$$\begin{aligned}
E\{e^{t(x-np)}\} &= E(e^{tx-t}) \\
&= e^{-tnp}\, E(e^{tx}) \\
&= e^{-tnp}\, M_{x}(t) \\
&= e^{-tnp}(q + pe^{t})^{n} \\
&= (qe^{-pt} + pe^{t-tp})^{n} \\
&= (qe^{-pt} + pe^{t-tp})^{n} \\
&= (qe^{-pt} + pe^{tq})^{n}
\end{aligned}$$

$$= \left[q\left\{1 - pt + \frac{p^{2}t^{2}}{2!} - \frac{p^{2}t^{3}}{3!} + \frac{p^{4}t^{4}}{4!} - \dots\right\} + p\left\{1 - tq + \frac{t^{2}q^{2}}{2!} + \frac{t^{2}q^{2}}{2!} + \frac{t^{3}q^{3}}{3!} + \dots\right\}\right]^{n}$$

$$= \left[\left\{1 + \frac{t^{2}}{2!}pq(p+q) + \frac{t^{3}}{3!}pq(q^{2} - p^{2}) + \frac{t^{4}}{4!}pq(q^{3} + p^{3}) + \dots\right\}\right]^{n}$$

$$= \left[\begin{array}{l} 1 + {}^{n}C_{1}\left\{\dfrac{t^{2}}{2!}pq + \dfrac{t^{3}}{3!}pq(q-p) + \dfrac{t^{4}}{4!}pq(1-3pq) + \dots\right\} \\ {}^{n}C_{2}\left\{\dfrac{t^{2}}{2!}pq + \dfrac{t^{3}}{3!}pq(q-p) + \dfrac{t^{4}}{4!}pq(1-3pq) + \dots\right\}^{2} + \dots \end{array}\right]$$

Since we have, $a^{3} + b^{3} = (a + b)(a^{2} - ab + b^{2})$; $p^{3} + q^{3} = (p + q)(p^{2} - pq + q^{2})$

$$= (1)(p + q)^{2} - 3(pq) = (1 - 3pq)$$

Now $\mu_{2} =$ coefficient of $\frac{t^{2}}{2!} = npq$; $\mu_{3} =$ coefficient of $\frac{t^{3}}{3!} = npq(q-p)$;

$\mu_{4} =$ coefficient of $\frac{t^{4}}{4!}$

$$= npq(1-3pq) + \frac{n!}{2!(n-2)!}6\frac{t^{4}}{4\times 3\times 2}p^{2}q^{2}$$

$$= npq(1-3pq) + \frac{n(n-1)(n-2)!}{2!(n-2)!}6p^{2}q^{2}$$

$$= npq(1 - 3pq) + 3n(n-1)p^{2}q^{2}$$

Additive Property of Binomial Distribution

Let $X \sim b(n_1, p_1)$ and $Y \sim b(n_2, p_2)$ be independent random variables.

Then $M_X(t) = (q_1 + p_1 e^t)^{n_1}; M_Y(t) = (q_2 + p_2 e^t)^{n_2}$

What is distribution of $X + Y$

We have $M_{X+Y}(t) = M_X(t)\, M_Y(t)$ [since X and Y are independent]

$= (q_1 + p_1 e^t)^{n_1}; (q_2 + p_2 e^t)^{n_2}$

Since it cannot be expressed in the form $(q + pe^t)^n$, from uniqueness theorem of m.g.f it follows that $X + Y$ is not a binomial variate. Hence, in general the sum of two independent binomial variates is not a binomial variate.

In other words, binomial distribution does not possess the additive or reproductive property.

However, if we take $p_1 = p_2 = p$, we get

$$M_{X+Y}(t) = (q + pe^t)^{n_1 + n_2}$$

which is the m.g.f of a binomial variate with parameters $(n1 + n2, p)$. Hence, by uniqueness theorem of m.g.f's $X + Y \sim b(n_1 + n_2, p)$. Thus the binomial distribution possesses the additive or reproductive property if $p_1 = p_2$.

Generalization: If X_i for all $i = 1, 2, 3, \ldots, k$ then their sum $\sum_{i=1}^{k} X_i \approx B\left(\sum_{i=1}^{k} n_i p\right)$.

Recurrence Relation for the probabilities of Binomial Distribution

(Fitting of Binomial Distribution)

We have $P(x + 1) = {}^nC_{x+1}\, p^{x+1}\, q^{n-x-1}$ and $P(x) = {}^nC_x p^x\, q^{n-x}$

$$\therefore \quad \frac{P(x+1)}{P(x)} = \frac{{}^nC_{x+1}p^{x+1}q^{n-x-1}}{{}^nC_x p^x\, q^{n-x}} = \frac{\dfrac{n!}{(x+1)!(n-x-1)!}}{\dfrac{n!}{x(n-x)!}}\,\frac{p}{q}$$

$$\frac{P(x+1)}{P(x)} = \frac{x!\,(n-x)\,(n-x-1)!\,p}{(x+1)\,x!\,(n-x-1)!\,q} = \left(\frac{n-x}{1+x}\right)\frac{p}{q}$$

$\therefore P(x + 1) = \left\{\left(\dfrac{n-x}{x+1}\right)\dfrac{p}{q}\right\} P(x)$ which is the required recurrence formula.

Example 2.8 It has been claimed that in 60% of solar heat installations the utility bills is reduced by least one-third. Accordingly what are the probabilities that utility bill will be reduced by at least one-third in (*i*) four of five installations (*ii*) at least four of five installations?

Solution: $n = 5, p = 0.6, q = 1 - p = 0.4$

(*i*) $b(4; 5, 0.6) = {}^5C_4\,(0.6)^4\,(0.4)^1 = 5(0.6)^4\,(0.4) = 0.2592$

(*ii*) at least 4 means 4 or 5

$$b(5;\,5,\,0.6) = {}^5C_5\,(0.6)^5\,(0.4)^0 = 0.0778$$

$\therefore$ Probability in at least four installations $= b(4;\,5,\,0.6) + b(5;\,5,\,0.6)$

$$= 0.2592 + 0.0778 = 0.337$$

Hence the solution.

Example 2.9 Ten coins are tossed simultaneously. Find the probability of getting at least seven heads?

Solution:

$$n = 10,\ p = P(H) = \frac{1}{2};\ q = 1 - p = \frac{1}{2}$$

$$P(X \geq 7) = P(X = 7) + P(X = 8) + P(X = 9)\ P(X = 10)$$

$$= {}^{10}C_7(1/2)^7\,(1/2)^3 + {}^{10}C_8\,(1/2)^8\,(1/2)^2 + {}^{10}C_9\,(1/2)^9\,(1/2)^1 + {}^{10}C_{10}\,(1/2)^{10}\,(1/2)^0$$

$$= \frac{1}{2^{10}}\left[{}^{10}C_7 + {}^{10}C_8 + {}^{10}C_9 + {}^{10}C_{10}\right] = \frac{1}{2^{10}}\left[{}^{10}C_3 + {}^{10}C_2 + {}^{10}C_1 + {}^{10}C_0\right]$$

$$= \frac{1}{2^{10}}\left[\frac{10.9.8}{1.2.3} + \frac{10.9}{1.2} + 10 + 1\right] = \frac{1}{2^{10}}[120 + 45 + 10 + 1]$$

$$= \frac{176}{2^{10}} = 0.172$$

Hence, the solution.

Example 2.10 If 3 of 20 tyres are defective and 4 of them are randomly chosen for inspection. What is the probability that only one of the defective tyres will be included?

Solution:

$$n = 4,\ p = \frac{3}{20},\ q = 1 - p = \frac{17}{20}$$

$$P(x = 1) = {}^4C_1\,(p)^1\,(q)^{(4-1)}$$

$$= 4 \cdot \frac{3}{20} \cdot \left(\frac{17}{20}\right)^3 = \frac{4.3.17^3}{20^4} = 0.368$$

Hence, the solution.

Chebyshev's Theorem: Let X be a random variable with mean μ and standard deviation σ then $P[(|x - \mu|) \geq k\sigma)] \leq \frac{1}{k^2}$.

Proof: Let $f(X)^2$ be the probability mass function of a random variable having mean μ and variance σ.

Now $\quad \sigma^2 = \Sigma_x\,(x - \mu^2)\,f(x)$

Let R_1 be the region in which $x \leq \mu - k\sigma$, R_2 the region in which $\mu - k\sigma < x < \mu + k\sigma$ and R_3 be the region in which $x \geq \mu + k\sigma$.

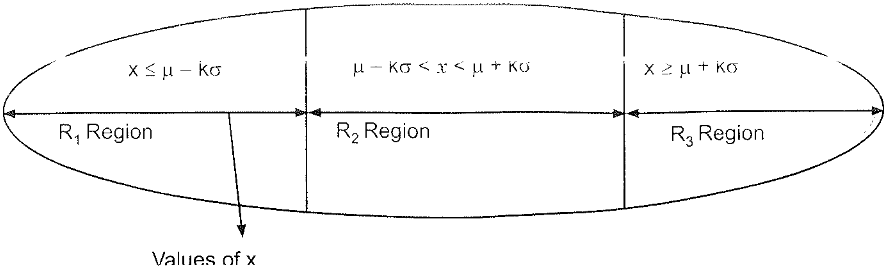


Fig. 2.3 Chebyshev's theorem

$$\therefore \quad \sigma^2 = \Sigma_{R_1}(x-\mu)^2 f(x) + \Sigma_{R_2}(x-\mu)^2 f(x) + \Sigma_{R_3}(x-\mu)^2 f(x)$$

Since $(x-\mu)^2 \geq 0$ i.e., non-negative, hence $\Sigma_{R_2}(x-\mu)^2 f(x) \geq 0$ also non-negative.

Hence $\sigma^2 \geq \Sigma_{R_1}(x-\mu)^2 f(x) + \Sigma_{R_3}(x-\mu)^2 f(x)$

In R_1 $\quad x \leq \mu - k\sigma$

$\Rightarrow \quad x - \mu \leq -k\sigma$

In R_3 $\quad x \geq \mu + k\sigma$

$\Rightarrow \quad x - \mu \geq k\sigma$

Therefore; $\quad |x-\mu| \geq k\sigma$

Hence both R_1 and R_3, $(x-\mu)^2 \geq k^2\sigma^2$

And $\quad \sigma^2 \geq \Sigma_{R_1} k^2\sigma^2 f(x) + \Sigma_{R_3} k^2\sigma^2 f(x)$

i.e., $\quad \sigma^2 \geq k^2\sigma^2 [\Sigma_{R_1} f(x) + \Sigma_{R_3} f(x)]$

$$\Rightarrow \quad \frac{1}{k^2} \geq [\Sigma_{R_1} f(x) + \Sigma_{R_3} f(x)]$$

Now $[\Sigma_{R_1} f(x) + \Sigma_{R_3} f(x)]$ represents the probability assigned to the region $R_1 \cup R_3$.

$$\therefore \quad \left[\sum_{R_1} f(x) + \sum_{R_3} f(x)\right] = P[|x-\mu| \geq k\sigma]$$

$$\therefore \quad \frac{1}{k^2} \geq P[|x-\mu| \geq k\sigma]$$

$$\Rightarrow \quad P[|x-\mu| \geq k\sigma] \leq \frac{1}{k^2}$$

This completes the proof of the theorem

Note: $P[|x-\mu| \geq k\sigma] \leq \dfrac{1}{k^2} \quad \Rightarrow \quad P[|x-\mu| < k\sigma] \leq 1 - \dfrac{1}{k^2}$

2.7.2 Poisson Distribution

A random variable X is said to follow Poisson distribution if its probability mass function is given by $f(x, \lambda) = \dfrac{\lambda^x}{x!} e^{-\lambda}$, for $x = 0, 1, 2, \ldots$

Poisson Approximation to Binomial Distribution Theorem

Statement: As $n \to \infty$ and $p \to 0$ so that $np = \lambda$ where λ is a finite non-zero constant then $b(x, n, p) \to \dfrac{\lambda^x}{x!} e^{-\lambda}$.

Proof: Let us consider $b(x, n, p)$ so that $b(x, n, p) = {}^nC_x\, p^x\, q^{n-x}$

$$= \frac{n(n-1)(n-2)\ldots(n-(x-1))}{x!} p^x q^{n-x}$$

and given $np = \lambda \Rightarrow p = \dfrac{\lambda}{n}$ also $q = 1 - p = 1 - \dfrac{\lambda}{n}$

$$= \frac{n(n-1)(n-2)\ldots(n-(x-1))}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{x!} \frac{n(n-1)(n-2)\ldots(n-(x-1))}{n^x} \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^x}$$

$$= \frac{\lambda^x}{x!}\left(\frac{n}{n}\right)\left(\frac{(n-1)}{n}\right)\left(\frac{(n-2)}{n}\right)\ldots\left(\frac{n-(x-1)}{n}\right)\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^x}$$

$$b(x, n, p) = {}^nC_x\, p^x\, q^{n-x}$$

$$= \frac{\lambda^x}{x!}\left(\frac{n}{n}\right)\left(\frac{(n-1)}{n}\right)\left(\frac{(n-2)}{n}\right)\ldots\left(\frac{n-(x-1)}{n}\right)\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^x}$$

Now as $n \to \infty$ $\dfrac{1}{n}, \dfrac{2}{n}, \ldots, \dfrac{x-1}{n} \to 0$

$$\left(1-\frac{\lambda}{n}\right)^{x} \to 1 \text{ and } \left(1-\frac{\lambda}{n}\right)^{n} = \left[\left(1-\frac{\lambda}{n}\right)^{-n/\lambda}\right]^{-\lambda} \to e^{-\lambda}$$

$$\therefore \qquad b(x, n, p) \to \frac{\lambda^{x}}{x!} e^{-\lambda}.$$

This completes the proof of the Poisson's Approximation to Binomial distribution theorem.

Note:

1. $e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^{x}}{x!} = \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!}$
2. Show that $\sum_{x=0}^{\infty} f(x, \lambda) = 1$

   For that consider $\sum_{x=0}^{\infty} f(x, \lambda) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x}}{x!} = e^{-\lambda} e^{\lambda} = 1$
3. $\lambda > 0$ is called the parameter of the Poisson Distribution.
4. $P(X = 0) = \frac{e^{-\lambda}\lambda^{0}}{0!} = e^{-\lambda}$

Applications of Poisson Distribution

Poisson distribution is applicable when n is very large and p is very small. Hence some of the applications of Poisson distribution are as follows:

1. Number of faulty blades produced by a reputed firm.
2. Number of deaths from a disease such as heart attack or cancer.
3. Number of telephone calls received at a particular telephone exchange.
4. Number of cars passing a crossing per minute.
5. Number of printing mistakes in a page of a book.

Mean and Variance of Poisson Distribution

$$\begin{aligned} \text{Mean } \mu &= E(X) \\ &= \sum_{x=0}^{\infty} xP(X = x) = \sum_{x=0}^{\infty} xf(x, \lambda) \\ &= \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^{x}}{x!} = \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{(x-1)!} \\ &= \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} e^{\lambda} \end{aligned}$$

$$\text{Mean} = \mu = \lambda$$

$$\begin{aligned}
E(X^2) &= \sum_{x=0}^{\infty} x^2 \, f(x,\lambda) \\
&= \sum_{x=0}^{\infty} [x(x-1)+x] \, f(x,\lambda) \\
&= \sum_{x=0}^{\infty} x(x-1) \, f(x,\lambda) + \sum_{x=0}^{\infty} x f(x,\lambda) \\
&= \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} + \lambda \\
&= \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-2)!} + \lambda \\
&= \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda
\end{aligned}$$

$$\therefore \quad E(X^2) = \lambda^2 \, e^{-\lambda} \, e^{\lambda + \lambda}$$

$$E(X^2) = \lambda^2 + \lambda$$

$$\text{Variance} = V(X) = \sigma^2 = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

$$\therefore \text{ variance} = \lambda$$

$$\text{Standard Deviation} = \text{S.D.} = \sqrt{Variance} = \sqrt{\lambda}$$

Note: In a Poisson distribution mean always equal to the variance.

Moment Generating Function of Poisson Distribution

$$\begin{aligned}
M_X(t) &= E[e^{tx}] \\
&= \sum_{x=0}^{\infty} e^{tx} f(x,\lambda) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{\lambda}\lambda^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t\lambda)^x}{x!} = e^{-\lambda} \, . \, e^{\lambda e^t} = e^{\lambda(e^t - 1)}
\end{aligned}$$

Additive Property of Poisson Variates

Theorem: If X and Y are two independent random Poisson variates with parameters λ and μ then $X + Y$ is also a Poisson variate with parameter $\lambda + \mu$.

Proof: Since X is a Poisson variate with parameter $\lambda \Rightarrow M_X(t) = e^{\lambda(e^t - 1)}$

Similarly, since Y is Poisson variate with parameter $\mu \Rightarrow M_Y(t) = e^{\mu(e^t - 1)}$

From the additive property of the moment generating function

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) \\ &= e^{\lambda(e^t - 1)} e^{\mu(e^t - 1)} \\ &= e^{(\lambda+\mu)(e^t - 1)} \end{aligned}$$

Which is the moment generating function of a Poisson variate with parameter $\lambda + \mu$.

$\therefore X + Y$ is also a Poisson variate with parameter $\lambda + \mu$.

2.7.3 Poisson Process

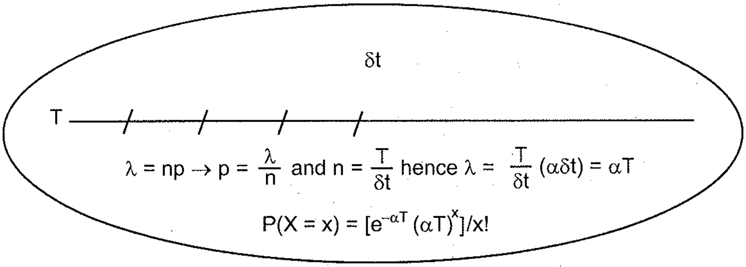

δt
T
λ = np → p = λ/n and n = T/δt hence λ = T/δt (αδt) = αT
P(X = x) = [e^{-αT} (αT)^x]/x!

Fig. 2.4 Poisson Process

Suppose we have to find the probability of x successes during a time interval T. Divide the time interval T into n equal parts of width δt. Therefore $T = n.\delta t$.

Here we make the following assumptions:

(a) The probability of success during an interval δt is given by $\alpha.\delta t$.

(b) The probability more than one success in a small interval δt is negligible.

(c) The probability of success in interval $(t, t + \delta t)$ is independent of the actual time t and also of all previous successes.

Here the assumptions of binomial distribution are satisfied and the probability $b(x, n, p)$ where $n = \frac{T}{\delta t}$ and $p = \alpha \cdot \delta t$.

$\therefore$ As $n \to \infty$ binomial distribution approaches to Poisson distribution and here parameter l is

$$\lambda = np \quad \to \quad p = \frac{\lambda}{n}$$

$$n = \frac{T}{\delta t} \text{ and } p = \alpha\delta t$$

Therefore; $$\lambda = \frac{T}{\delta t}(\alpha\delta t) = \alpha T$$

$$\boxed{\therefore P(X = x) = e^{-\alpha T}(\alpha T)^x/x!}$$

Example 2.11 If the probability that an individual suffers a bad reaction due to a certain injection is 0.001, determine the probability that out of 2000 individuals (*i*) exactly 3 (*ii*) more than 2 individuals will suffer a bad reaction?

Solution: Given $p = 0.001; \quad n = 2000; \quad \lambda = np = 2$

(i) To find $P(\text{Exactly } 3) = P(X = 3)$

$$= \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-2}2^3}{3!} = \frac{4}{3e^2}$$

$$= 0.1804 \text{ since } e = 2.086,\ 2 < e < 3$$

(ii) $P(\text{more than 2 individuals}) = P(X > 2)$

$$= 1 - P(X \le 2) = 1 - [P(X = 0) + P(x = 1) + P(x = 2)]$$

$$= 1 - \left[\frac{e^{-\lambda}\lambda^0}{0!} + \frac{e^{-\lambda}\lambda^1}{1!} + \frac{e^{-\lambda}\lambda^2}{2!}\right]$$

$$= 1 - e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2}\right] = 1 - 5e^{-2} = 0.323.$$

Hence, the solution.

Example 2.12 A manufacturer of cotter pins knows that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality?

Solution: We are given $n = 100$, p = probability of defective pins = 5% = 0.05

And λ = mean number of defective pins in a box of 100 = $np = 100 \times 0.05 = 5$

Since p is small, we may use Poisson distribution probability of 'x' defective pins in a box of 100 is $P(X = x) = \dfrac{e^{-\lambda}\lambda^x}{x!} = \dfrac{e^{-5}5^x}{x!}$ for all $x = 0, 1, 2, \ldots$

Probability that a box will fail to meet the guaranteed quality is

$$P(X > 10) = 1 - P(X \le 10)$$

$$= 1 - \sum_{x=0}^{10}\frac{e^{-5}5^x}{x!} = 1 - e^{-5}\sum_{x=0}^{10}\frac{5^x}{x!}$$

Hence the solution.

Example 2.13 10% of the bolts produced by a certain machine turn out to be defective. Find the probability that in a sample of 10 tools selected at random exactly two will be defective using (*i*) binomial distribution (*ii*) Poisson distribution and comment upon the result?

Solution: Given $p = \frac{10}{100} = 0.1, n = 10, \lambda = np = 1$

(*i*) Using binomial distribution

Let $q = 1 - p = 1 - 0.1 = 0.9$

$$P(X = 2) = {}^{10}C_2\, p^2\, q^{(n-2)} = \frac{10 \times 9}{1 \times 2}(0.1)^2\,(0.9)^8 = 0.194$$

(*ii*) Using Poisson distribution

$$P(X = 2) = \frac{e^{-\lambda}\lambda^2}{2!} = \frac{e^{-1}1^2}{2} = \frac{1}{2e} = 0.184$$

Comment: There is a difference between the two probabilities because of the fact that Poisson distribution (P.D.) is an approximation to binomial distribution (B.D.) and it is applicable for large *n*. Hence the solution.

Example 2.14 A hospital switch board receives an average of 4 emergency calls in a 10 minutes interval. What is the probability that (i) there are at the most 2 emergency calls and (ii) there are exactly 3 emergency calls in a 10 minutes interval?

Solution: Given, $\lambda = 4,$

(*i*)
$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$$
$$= \frac{e^{-\lambda}\lambda^0}{0!} + \frac{e^{-\lambda}\lambda^1}{1!} + \frac{e^{-\lambda}\lambda^2}{2!}$$
$$= e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2}\right]$$
$$= e^{-4}\,[1 + 4 + 8] = 13\, e^{-4} = 0.238.$$

(*ii*)
$$P(X = 3) = \frac{e^{-\lambda}\lambda^3}{3!} = \frac{e^{-4}4^3}{3!} = 0.195$$ Hence, the solution.

2.7.4 Normal Distribution (N.Dn)

Normal distribution is also a continuous distribution. A random variable X is said to follow normal distribution ($N.D^n$) with mean μ and variance σ^2 if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \qquad -\infty < x < \infty;\ -\infty < \mu < \infty;\ \sigma > 0$$

$$= 0 \qquad \text{otherwise.}$$

The corresponding distribution function is $f(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} dt$

Let $Z = \frac{x-\mu}{\sigma}$ then the mean of Z is 0 and the variance is 1.

The corresponding probability density function is $\varphi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(z)^2}{2}}, -\infty < z < \infty$

Z is called standard normal variate.

Notation:

1. $X \sim N(\mu, \sigma^2)$ denotes that X is a normal variate with mean μ and variance σ^2.
2. $Z \sim N(0, 1)$ denotes that z is a standard normal variate with mean 0 and variance 1.

Features of Normal Distribution curve

The graph of $f(x)$ is a bell shaped curve extending from – ∞ to ∞ with its peak at *m*.

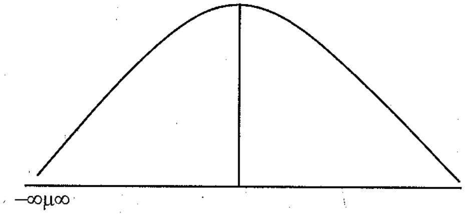


Fig. 2.5 Gaussian distribution φ(z)

Note:

1. The mode of normal distribution is μ.
2. The median of normal distribution is also μ. Hence, for a normal distribution the mean, median and mode coincide.

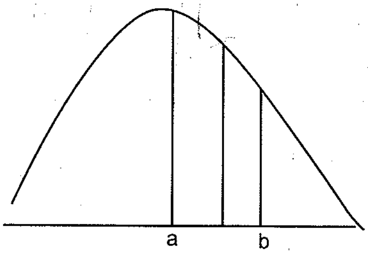


Fig. 2.6 Area under the normal curve

The area under the normal curve between the ordinates $x = a$ and $x = b$ gives the probability that the random variable X lies between a and b.

$$\therefore \quad P(a < X < b) = \int_a^b f(x)dx = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Put $\quad Z = \dfrac{x-\mu}{\sigma}$

So $\quad dz = \dfrac{dx}{\sigma} \Rightarrow dx = \sigma \, . \, dz$

When $\quad x = a, z = \dfrac{a-\mu}{\sigma} = c$ (say)

When $\quad x = b, z = \dfrac{b-\mu}{\sigma} = d$ (say)

$$\therefore \quad P(a < X < b) = \int_c^d \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{z^2}{2}} \sigma \, . \, dz = \int_c^d \varphi(z)dz = P(c < z < d)$$

So,

1. $\int_{-\infty}^{\infty} \varphi(z)\, dz = 1$
2. $\int_{-\infty}^{\infty} \varphi(z)\, dz = 2\int_0^{\infty} \varphi(z)\, dz = 2 \times 0.5 = 1$
3. $\int_{-a}^{a} \varphi(z)\, dz = 2\int_0^{a} \varphi(z)\, dz$

The $\int_0^z \varphi(z)dz$ are available in the CDF table of standard normal distribution.

Example 2.15 Fit a Poisson distribution to the following data and calculate the theoretical frequencies:

x	0	1	2	3	4
y	46	38	22	9	1

Solution: Mean $\mu = E(X) = \lambda$ and Variance $V(X) = \sigma^2 = E(X)^2 - [E(X)]^2$

x_i	f_i	$f_i x_i$	x_i^2	$f_i x_i^2$
0	46	0	0	0
1	38	38	1	38
2	22	44	4	88
3	9	27	9	81
4	1	4	16	16

	$\sum_{x=0}^{4} f_i = 116$	$\sum_{x=0}^{4} f_i x_i = 113$		$\sum_{x=0}^{4} f_i x_i^2 = 223$

$$\text{Mean} = \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{113}{116} = 0.974;$$

$$\text{Variance} = \frac{\Sigma f_i x_i^2}{\Sigma f_i} - (\bar{x})^2 = \frac{223}{116} - (0.974)^2 = 1.9224 - 0.9487 = 0.974$$

$$\therefore \quad \text{Mean} = \text{Variance} = \lambda = 0.974$$

The theoretical frequencies are $f(x) = N.P(X = x)$

$$f(0) = 116\,P(X = 0) = 116 \times e^{-0.974} = 44$$

$$f(1) = 116\,P(X = 1) = 116 \times e^{-0.974}\,0.974 = 42$$

$$f(2) = 116\,P(X = 2) = \frac{116 \times e^{-0.974}(0.974)^2}{2!} = 21$$

$$f(3) = 116\,P(X = 3) = \frac{116 \times e^{-0.974}(0.974)^3}{3!} = 7$$

$$f(4) = 116 \times P(X = 4)$$

$$= 116 - \{f(0) + f(1) + f(2) + f(3)\} = 116 - 114 = 2$$

Hence the solution.

Example 2.16 If a bank receives on an average 6 bad cheques per day, what are the probabilities that it will receive (i) four bad cheques on any given day (ii) 10 bad cheques on any two consecutive days.

Solution: Let

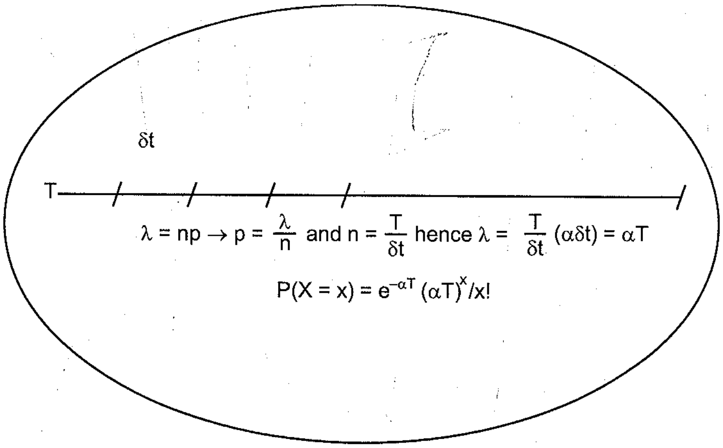

δt
T
$\lambda = np \rightarrow p = \frac{\lambda}{n}$ and $n = \frac{T}{\delta t}$ hence $\lambda = \frac{T}{\delta t}(\alpha\delta t) = \alpha T$
$P(X = x) = e^{-\alpha T}(\alpha T)^x/x!$

Fig. 2.7

(i) $\alpha = 6, T = 1$ and $\therefore \lambda = \alpha T = 6$

$$f(4, 6) = \frac{e^{-6}6^4}{4!} = 0.1339$$

(i) $\alpha = 6, T = 2$ and $\therefore \lambda = \alpha T = 12$

$$f(10; 12) = \frac{e^{-\lambda}\lambda^{10}}{10!} = \frac{e^{-12}12^{10}}{10!} = 0.105$$

Hence, the solution.

PROBLEMS

1. A random variable X has the following probability distribution:

x	0	1	2	3	4	5	6	7	8
P(x)	$2k$	k	$2k$	$4k$	$5k$	k^2	$2k^2$	$7k^2$	$6k$

Determine

(i) k

(ii) $P(X < 3)$

(iii) $P(X > 6)$

(iv) $P(0 < X < 3)$

2. Let X and Y be jointly continuous random variables with joint PDF

$$f_{X,Y}(x, y) = \begin{cases} cx + 1 & x, y \geq 0, x + y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Show the range of (X, Y), R_{XY}, in the x-y plane.

Find the constant c.

Find the marginal PDFs $f_X(x)$ and $f_Y(y)$

Find $P(y < 2x^2)$

3. Let X be a continuous random variable with the following PDF

$$f_X(x) = \begin{cases} ce^{-4x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where c is a positive constant.

(a) Find c.

(b) Find the CDF of X, $F_X(x)$

(c) Find $P(2 < X < 5)$

(d) Find $E[X]$

4. Let $X \sim N(3, 9)$ and $Y = 5 - X$

(a) Find $P(X > 2)$.

(b) Find $P(-1 < Y < 3)$

(c) Find $P(X > 4 | Y < 2)$

5. A car hire firm has two cars which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with mean 1.5. calculate the proportion of days. (*i*) On which there is no demand. (*ii*) On which demand is refused ($e^{-5} = 0.2231$)?

6. X is a normal variate with mean 30 and standard deviation 5. Find the probabilities that (*i*) $26 \leq X \leq 40$, (*ii*) $X \geq 45$.

7. In a normal distribution ($N.D^n$) 31% of the items are under 45 and 8% are over 63. Find the mean and variance of the distribution.

CHAPTER

3

SAMPLING DISTRIBUTION

Chapter Outline

• Introduction • Some Common Terms Used in Sampling Distributions • Sampling Distributions

3.1 INTRODUCTION

The field of statistics encompasses the processes of collecting, presenting, analyzing, and utilizing data to inform decision-making and problem-solving. The primary goal of any statistical inquiry is to derive meaningful conclusions about a group of subjects under examination, referred to as the population.

Given the challenges or impracticalities associated with scrutinizing en entire population, the concept of investigating only a representative subset, known as a sample, emerges. This exploration seeks to draw inferences about the entire population by leveraging insights gained from the sample – a method termed statistical inference. The act of selecting samples is denoted as sampling, and a sample is deemed a faithful representation of the population when a probabilistic sampling method is employed. Notably, random sampling stands out as the most pivotal among probabilistic sampling techniques, ensuring that each population member has an equal likelihood of inclusion in the sample.

Samples play a crucial role in making inferences about populations by estimating key parameters such as mean (μ), standard deviation (σ), and others. The estimation of population parameters relies on the examination of pertinent statistical quantities derived from a sample – referred to as sample statistics. The term 'statistic' is commonly used to denote either the random variable or its value, with the specific meaning discerned from the surrounding context.

Let us consider all possible samples of a population and calculate a statistic for instance sample mean. Then the set of all such b\values, one for each sample, is called the sampling distribution of the statistic.

Now we can compute the statistics mean variance, etc., for this sampling distribution.

In most statistic problems, it is necessary to use the information from sample to draw inferences about the population.

3.2 SOME COMMON TERMS USED IN SAMPLING DISTRIBUTIONS

Population

The population in a statistical study is the set or collection or totality of observations about which inferences are to be drawn. Thus the population consists of sets of numbers, measurements or observations. Population size N is the number of objects or observations in the population.

Population is said to be finite or infinite depending on the size N being finite or infinite. Since it is impracticable to examine the entire population, a finite subset of the population known as sample is studied. Sample size n is the number of objects or observations in the sample.

Example 3.1 (i) Engineering graduate students in A.P. (Population), Engineering graduate students of a college (sample).

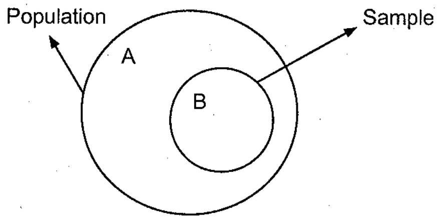


Fig. 3.1 Example 3.1

Example 3.2 Total production of items in a month (Population), Total production of items in one day (Sample).

Example 3.3 Budget of India (Population), Budget of A.P. (Sample), budget of a district (sub sample).

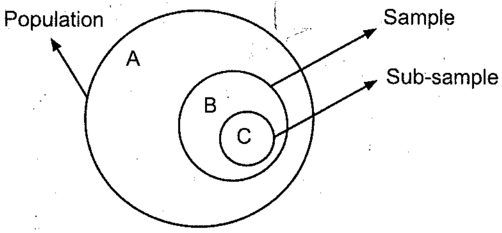


Fig. 3.2 Population, Sample and Sub-sample

Population parameter: A statistical measure or constant obtained from the population is called a population parameter. For example, population mean (μ), population variance (σ^2).

(Sample) statistic: A statistical measurement computed from sample observations is called a (sample) statistic. For example, sample mean $(\bar{x})$, sample variance (σ^2) clearly, parameters are to population while statistics are to sample.

μ, σ, p represent the population mean, population standard deviation, population proportion, similarly $\bar{x}$, s, p denote sample mean, sample standard deviation (s.s.d.), sample proportion.

Note: The samples must be a true or good representative of the population, sampling should be random or probabilistic.

Sampling: The process of drawing or obtaining samples is called sampling.

Large sampling: If $n \geq 30$, then the sampling is known as large sampling.

Small sampling: If $n < 30$, then the sampling is known as small or exact sampling.

Note: The simplest and most commonly used type of probabilistic sampling is the random sampling.

Random Sampling: Each member of the population has equal chances or probability of being included in the sample. The sample obtained by this method is termed as a random sample.

Finite Population: Population may be finite or infinite. If the number of items or observations consisting the population is fixed and limited, it is called as finite population.

Example 3.4 The workers in a factory, student in a college etc.,

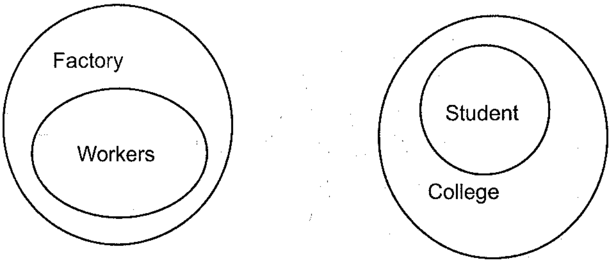


Fig. 3.3 Example 3.4

Infinite Population: If the number of items or observations consisting the population is infinite (not fixed and not limited), it is called as finite population.

Example: The population of all real numbers lying between 0 and 1. The population of stars or astral bodies in the sky.

Sampling with replacement: If the items are selected or drawn one by one such a way that an item drawn at a time is replaced back to the population before the next or subsequent draw, it is known as (random) sampling with replacement.

In this type of sampling from a population of size N, the probability of a selection of a unit at each draw remains $\frac{1}{N}$. Thus sampling from finite population with replacement can be considered theoretically as sampling from infinite population. In this, N_n samples will be drawn.

In Sampling without replacement: An item of the population cannot be chosen for more than once, as it is not replaced. In this N_{C_n} samples will be drawn. Hence the probability of drawing a unit from a population of N items at r^{th} draw is $\frac{1}{N-r+1}$.

Statistic is a real-valued function of the random sample. So it is a function of one or more random variables not involving any unknown parameter. Thus statistic is a function of samples observations only and is itself a random variable. Hence, a statistic must have a probability distribution.

Sample mean: Let $x_1, x_2, x_3, ..., x_n$ be a random, sample of size n from a population.

$$\text{Then sample mean} = (\overline{x}) = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Sample Variance: Then sample variance $= s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$.

Sample standard deviation is the positive square root of sample variance. Sample mean and sample variance are two important statistics which are statistical measures of a random sample of size n.

3.3 SAMPLING DISTRIBUTIONS

Let us consider all possible samples of size n, from a finite population of size N. Then the total number of all possible samples of size n, which can be drawn from the population is $N_{C_n} = m$.

Compute a statistic θ [such as mean, variance/s.d, proportion] for each of these sample using the sample data $x_1, x_2, x_3, ..., x_n$ by $\theta = \theta(x_1, x_2, x_3, ..., x_n)$.

Sample number	1	2	3	...	m
Statistic θ	θ_1	θ_2	θ_3	...	θ_m

Sampling distribution of the statistic θ is the set of values $\{\theta_1, \theta_2, \theta_3, ..., \theta_m\}$ of the statistic θ obtained, one for each sample. Thus sampling distribution describes how a statistic θ will vary from one sample to the other of the same size. Although all the m samples are drawn from the given population, the items included in different samples are different.

If the statistic θ is mean, then the corresponding distribution of the statistic is known as sampling distribution of means, thus if θ is variance, proportion etc., the corresponding distribution is known as sampling distribution of variances, sampling distribution of proportions etc.,

Then Mean of sampling distribution of $\theta = (\bar{\theta}) = \dfrac{\sum_{i=1}^{m} \theta_i}{n}$

And Variance of sampling distribution of $\theta = \dfrac{\sum_{i=1}^{m} (\theta_i - \bar{\theta})^2}{m}$.

Similarly, we can have mean of sampling distribution of means, variance of sampling distribution of means, variance of the sampling distribution of variances etc.

Standarad Error

The standard deviation of the sampling distribution of a statistic is known as Standard Error (SE). The standard error gives some idea about the precision of the estimate of the parameters. As the sample size n increases, SE decreases. SE plays a very important role in large sample decision theory and forms the basis for hypothesis testing.

Sampling distribution of a statistic enables us to get information about the corresponding population parameter.

Degrees of Freedom **(ν)**

The number of degrees of freedom usually denoted by greek alphabet ν, is a positive integer equals to $n - k$ where n is the number of independent observations of the random sample and k is the number of population parameters which are calculated using the sample data. The degrees of freedom $\nu = n - k$ is the difference between n the sample size and k the number of independent constaints imposed on the observations in the sample.

3.3.1 The Sampling Distribution of the Mean (σ known)

To answer any questions related to sampling distribution of the mean $(\overline{X})$ we need to consider a random sample of n observations and determine the value $\bar{x}$ for each sample, then by various values of $\overline{X}$, it may be possible to get an idea of the nature of the sampling distribution. Aslo we have to consider the following theorem for the mean $\mu_{\bar{x}}$ and the variance $\sigma_{\bar{x}}^2$ of sampling distribution of the mean $(\overline{X})$.

Theorem: If a random sample of size n is taken from a population having the mean μ and the variance σ^2, then $(\overline{X})$ is a random variable whose distribution has the mean μ.

Proof: For samples from infinite population the variance of this distribution is $\dfrac{\sigma^2}{n}$.

For samples from finite population the variance of this distribution is $\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$.

By above statement, population is infinite then sampling with replacement

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

And when the population is finite, size N (sampling without replacement)

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}$$

Note: The factor $\left(\frac{N-n}{N-1}\right)$ is known as finite population correction factor.

In sampling with replacement, we will have N^n samples each with probability $\frac{1}{N^n}$.

In sampling without replacement we will have N_{C_n} samples each with probability $\frac{1}{N_{C_n}}$.

Note: The factor $\left(\frac{N-n}{N-1}\right)$ can be neglected if N is too large compared to the sample size n.

3.3.2 Central limit theorem

Whenever n is large, the sampling distribution of $\bar{X}$ approximately normal with mean μ and variance —— regardless of the form of the parent population distribution, as the following theorem states [without proof]

Theorem: If $\bar{X}$ is the mean of a random sample of size n drawn from a population with mean μ and finite variance σ^2 then the standardized sample mean $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is a random variable whose distribution function approaches that of the standard normal distribution $N(0, 1)$ as $n \to \infty$.

Normal distribution provides a good approximation to the sampling distribution for almost all the populations for $n \geq 30$.

For $n < 30$ small samples, sampling distribution of $\bar{X}$ is normally distributed, provided sampling is from normal population.

3.3.3 Sampling Distribution of Proportions

Suppose that a population is infinite and that the probability of occurrence of an event called its success is p, while the probability of non-occurrence of the event is $q = 1 - p$. Consider all possible samples of size N drawn from tis population, and for each sample compute the proportion p of successes. Then, we can have a sampling distribution of proportions whose mean samples deviation σ_p are given by

$$\mu_p = p$$

And

$$\sigma_p^2 = \frac{p(1-p)}{n} = \frac{pq}{n}$$

While population is binomially distributed, the sampling distribution of proportion is normally distributed whenever n is large $n \geq 30$. Above equations are also valid for a finite population in which sampling is with replacement.

For finite population sampling without replacement of size n

$$\mu_p = p \quad \text{and} \quad \sigma_p^2 = \frac{pq}{n}\left(\frac{N-n}{N-1}\right)$$

3.3.4 Sampling distributions of differences and sums

Let μ_{s_1} and σ_{s_1} be the mean and standard deviation of a sampling distribution of statistic s_1 obtained by calculating s_1 for all possible samples of size n_1 drawn from population 1. This yields a sampling distribution of the statistic s_1.

In an analogous manner, μ_{s_2} and σ_{s_2} be the mean and standard deviation of sampling distribution of statistic s_2 obtained by calculating s_2 for all possible samples of size n_2 drawn from another different population 2.

Now we can have a distribution of differences $s_1 - s_2$, called the sampling distribution of differences of the statistics, from the two population 1 and 2. Then the mean $\mu_{s_1 - s_2}$ and the standard deviation $\sigma_{s_1 - s_2}$ the sampling distribution of differences are given by

$$\mu_{s_1 - s_2} = \mu_{s_1} - \mu_{s_2}$$

and

$$\sigma_{s_1 - s_2} = \sqrt{\left(\sigma_{s_1}^2 + \sigma_{s_2}^2\right)}$$

provided the samples are independent.

Sampling distribution of sum of statistics has mean

$$\mu_{s_1 + s_2} = \mu_{s_1} + \mu_{s_2}$$

and

$$\sigma_{s_1 + s_2} = \sqrt{\left(\sigma_{s_1}^2 + \sigma_{s_2}^2\right)}$$

provided the samples are independent.

For infinite population the sampling distribution of the differences of means has mean $\mu_{\overline{X}_1 - \overline{X}_2}$ and standard deviation $\sigma_{\overline{X}_1 - \overline{X}_2}$ given by

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_{(\overline{X}_1)} - \mu_{(\overline{X}_2)} = \mu_1 - \mu_2$$

and $$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

For infinite population the sampling distribution of the differences of means has mean $\mu_{\overline{X}_1 + \overline{X}_2}$ and standard deviation $\sigma_{\overline{X}_1 + \overline{X}_2}$ given by

$$\mu_{\overline{X}_1 + \overline{X}_2} = \mu_{(\overline{X}_1)} + \mu_{(\overline{X}_2)} = \mu_1 + \mu_2$$

and $$\sigma_{\overline{X}_1 + \overline{X}_2} = \sqrt{\sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

3.3.5 Sampling Distribution of Mean σ Unknown: *t*-distribution

To estimate or infer on a population mean or the difference between two was assumed that the population standard deviation σ is known. When σ is unknown, for large $n \geq 30$, σ can be replaced by the sample standard deviation s, calculated using the sample mean $\overline{x}$ by the formula $= s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$.

For small sample of size $n < 30$ the unknown σ can be substituted by s, provided we make an assumption that the sample is drawn from a normal population.

Let $\overline{x}$ be the mean of a random sample of size n drawn from a normal population with mean μ and variance σ^2 then $t = \dfrac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$ is a random variable having the t-distribution with $\nu = n - 1$ degrees of freedom. Where $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$.

This result is more general than previous theorem CLT in the sense that it does not require knowledge of σ: on the other hand, it is less general than the previous theorem CLT in the sense that it requires the assumption of normal population.

Thus for all small samples $n < 30$ and with σ unknown a statistic for inference on population mean μ is $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$. With the underlying assumption of sampling from normal population.

The t-distribution curve is symmetric about the mean 0, bell shaped and asymptotic on both sides of horizontal t-axis.

Thus t-distribution curve is similar to normal curve. The variance for the t-distribution is more than 1 as it depends on the parameter $\nu = n - 1$ degrees of freedom but it approaches 1 as $n \to \infty$. In essence, as $\nu = (n - 1) \to \infty$, t-distribution tends to the standard normal distribution. Clearly for $n \geq 30$, standard normal distribution provides a good approximation to the t-distribution.

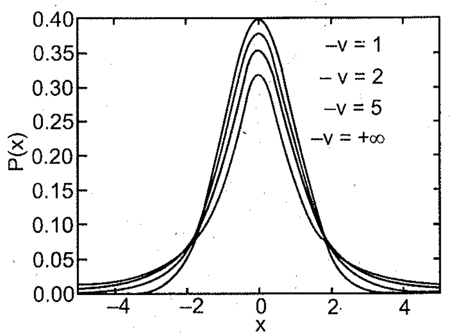


Fig. 3.4 ***t*-distribution**

Critical values of t-distribution is denote by t_α, which is such that the area under the curve to the right of t_α equals to α. Since the t-distribution is symmetric, it follows that $t_{1-\alpha} = -t_\alpha$, i.e., the t-value leaving an area of $1 - \alpha$ to the right and therefore an area α to its left, is equal to the negative t-value which leaves an area α in the right tail of the distribution.

Please observe critical values of t_α for values of the parameter ν. In tables the left-hand column contains values of ν, the column headings are area α in the right hand tail of the t-distribution, the entries are values of t_α.

3.3.6 χ^2 – Distribution

χ^2-(chi-squared) distribution is a continuous probability distribution of a c.r.v. X with probability density function given by

$$f(x) = \begin{cases} \dfrac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Where v is a +ve integer is the only single parameter of the distribution, also known as degrees of freedom.

Properties of χ^2 – Distribution

(*i*) χ^2 – Distribution curve is not symmetrical, lies entirely in the first quadrant. And hence not a normal curve, since χ^2 varies from 0 to ∞.

(*ii*) It depends only on the degrees of freedom ν.

(*iii*) If χ_1^2 and χ_2^2 are two independent distributions with ν_1, ν_2 degrees of freedom then $\chi_1^2 + \chi_2^2$ will be chi-squared distributions with $\nu_1 + \nu_2$ degrees of freedom – *i.e.*, it is additive.

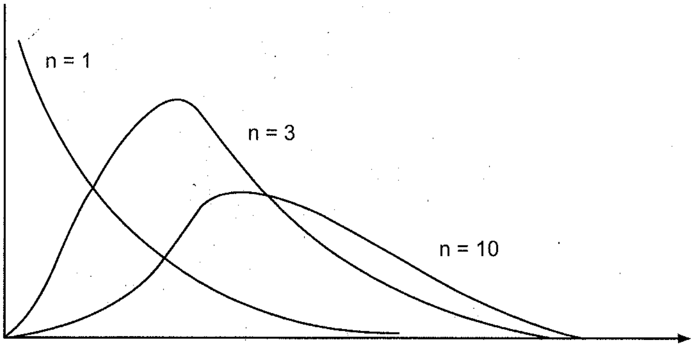


Fig. 3.5 χ^2 **– Distribution**

The mean (μ) of a chi-square distribution with ν degrees of freedom is ν. The variance (σ^2) is 2ν.

Applications: The chi-square distribution is commonly used in hypothesis testing, particularly in tests related to the comparison of observed and expected frequencies in categorical data. It is used in the construction of confidence intervals for the variance of a normal distribution.

A chi-square distribution table is a tabulated set of values that helps determine critical values and probabilities associated with the chi-square distribution. To use a chi-square distribution table, follow these steps:

For Critical Values:

1. **Identify Degrees of Freedom (ν):**
   - Determine the degrees of freedom (ν) associated with your chi-square distribution.
2. **Determine Significance Level (α):**
   - Choose a significance level (α) for your test. Common choices include 0.05, 0.01, or others depending on the specific requirements of your statistical test.

3. Look Up Critical Value:

- Locate the intersection of the row corresponding to the degrees of freedom (ν) and the column corresponding to the chosen significance level (α). The value in that cell is the critical value.

For Probabilities:

1. Identify Degrees of Freedom (ν):

- Determine the degrees of freedom (ν) associated with your chi-square distribution.

2. Determine Chi-Square Statistic (χ^2):

- Identify the chi-square statistic (χ^2) for which you want to find the probability.

3. Look Up Probability:

- Locate the intersection of the row corresponding to the degrees of freedom (ν) and the column corresponding to the chi-square statistic (χ^2). The value in that cell is the cumulative probability.

3.3.7 Sampling distribution of Variance s^2

From the earlier discussions, the sample mean is used to estimate the population mean. Similarly, the sample variance is used to estimate the population variance (σ^2). The sample variance is usually denoted by s^2 and is given by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

If s^2 is the variance of a random sample of size n from a normal population having the variance σ^2 then $\chi^2 = \frac{(n-1)\,s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sigma}$ is a random variable having the χ^2-distribution with $\nu = n - 1$ d of.

Exactly 95% of χ^2-distribution lies between $\chi^2_{0.975}$ and $\chi^2_{0.025}$ when σ^2 is too small. χ^2-value falls to the right of $\chi^2_{0.025}$ and when σ^2 is too large, χ^2 falls to the left of $\chi^2_{0.975}$. thus when σ^2 is correct χ^2-value fall s to the left of $\chi^{20.975}$ or to the right of $\chi^2_{0.025}$.

Table 3.1 Critical region for testing : $H_0 : \sigma^2 = \sigma_0^2$

Alternate hpothesis	Reject H_0 if
$\sigma^2 < \sigma_0^2$	$\chi^2 \leq \chi^2_{1-\alpha}$
$\sigma^2 > \sigma_0^2$	$\chi^2 \geq \chi^2_{1-\alpha}$
$\sigma^2 \neq \sigma_0^2$	$\chi^2 \leq \chi^2_{(1-\alpha)/2}$

3.3.8 *F*-Distribution (sampling distribution of the ratio of two sample variances)

Another important continuous probability distribution which plays an important role in connection with sampling from normal population is the *F*-distribution.

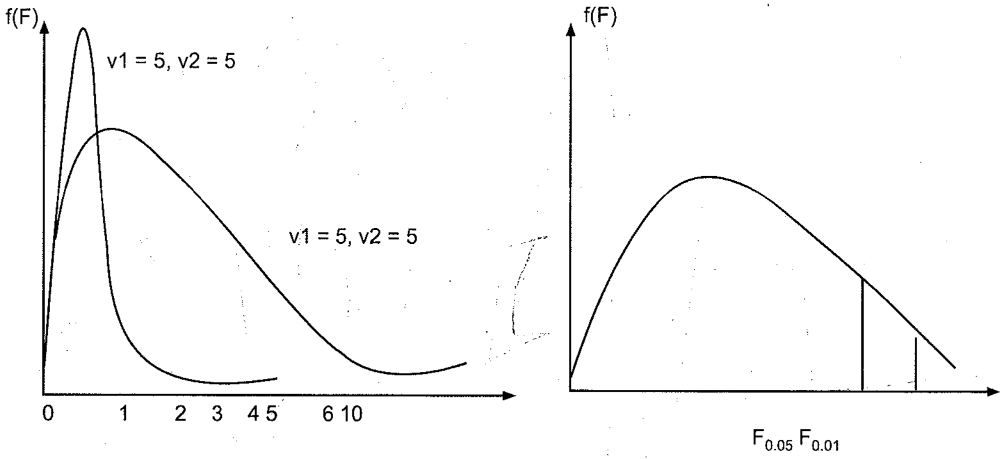
If s_1^2 and s_2^2 are the variances of independent random samples of size n_1 and n_2 from normal populations with variances σ_1^2 and σ_2^2.

To determine whether the two samples come from two populations having equal variances, consider the sampling distribution of the ratio of the variances of the two independent random samples defined by $F = \dfrac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} = \dfrac{\sigma_2^2\, s_1^2}{\sigma_1^2\, s_2^2}$ which follows *F*-distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

Uses: *F*-distribution can be used for testing the quality of several population means, comparing sample variances, and analysis of variance completely depends on *F*-distribution.

Under the hypothesis that two normal populations have the same variance: $\sigma_1^2 = \sigma_2^2$, we have $F = \dfrac{s_1^2}{s_2^2}$.

F determines whether the ratio of two sample variances s_1 and s_2 is too small or too large. When *F* is close to 1, the two sample variances s_1 and s_2 are almost same. *F* is always a positive number whenever the larger sample variance as the numerator.



Probability density function of several F distributions

Tabulated values of F

Fig. 3.6 *F*-Distribution

Properties of F-distribution

(*i*) *F*-distribution curve lies entirely in first quadrant.

(*ii*) The *F*-curve depends not only on the two parameters ν_1 and ν_2 but also on the order in which they are stated.

(*iii*) $F_{1-\alpha}(\nu_1, \nu_2) = \dfrac{1}{F_\alpha(\nu_2, \nu_1)}$ where $F_\alpha(\nu_2, \nu_1)$ is the value of *F* with ν_1 and ν_2 degrees of Freedom such that the area under the *F*-distribution curve to the right of F_α is α.

Note:

Table 3.2 Critical regions for testing the null hypothesis: $\sigma_1^2 = \sigma_2^2$

Alternate hypothesis	**Test statistic**	**Reject H_0 if**
$\sigma_1^2 < \sigma_2^2$	$F = \dfrac{s_1^2}{s_2^2}$	$F > F_\alpha(n_2 - 1, n_1 - 1)$
$\sigma_1^2 > \sigma_2^2$	$F = \dfrac{s_1^2}{s_2^2}$	$F > F\alpha(n_1 - 1, n_2 - 1)$
$\sigma_1^2 \neq \sigma_2^2$	$F = \dfrac{s_M^2}{s_m^2},\ s_M^2 > s_m^2$	$F > F_{\alpha/2}(n_M - 1, n_m - 1)$

Example 3.1 Find the value of the finite population correction factor for (*i*) $n = 10$ and $N = 1000$ (*ii*) $n = 100$ and $N = 1000$?

Solution: (*i*)

$$\frac{N-n}{N-1} = \frac{1000-10}{1000-1} = \frac{990}{999} = 0.991$$

(*ii*)

$$\frac{N-n}{N-1} = \frac{1000-100}{1000-1} = \frac{900}{999} = 0.900$$

Hence the solution.

Example 3.2 A random sample of size 2 is drawn from the population 3, 4, 5. Find (*i*) population mean (*ii*) Population SD (*iii*) Sampling distribution (SD) of means (*iv*) the mean of SD of means (*v*) SD of SD means?

Solution: (*i*) Population mean $= \mu = \dfrac{3+4+5}{3} = 4$

(*ii*) s.d. of population $= \sigma = \sqrt{\dfrac{(8-4)^2 + (4-4)^2 + (5-4)^2}{3}}$

$$= \sqrt{\frac{2}{3}} = \sqrt{0.6666} = 0.8164$$

(*iii*) sampling with replacement (infinite population): The total number of samples with replacement is $N^n = 3^2 = 9$ here N = population size and n = sample size. Listing all possible samples of size 2 from population 3, 4, 5 with replacement, we get 9 samples as below:

(3, 3) (3, 4) (3, 5)
(4, 3) (4, 4) (4, 5)
(5, 3) (5, 4) (5, 5)

Now compute the statistic the arithmetic mean for each of these 9 samples the set of 9 samples means $\overline{X}$, gives rise to the distribution of means of the sample known as sampling distribution of means

3 3.5 4
3.5 4 4.5
4 4.5 5

This sampling distribution of means can also be arranged in the form of frequency distribution

Sample mean $\overline{X}_i$	3	3.5	4	4.5	5
Frequency f_i	1	2	3	2	1

(*iv*) Mean of the sampling distribution of means

$$= \mu_{\overline{X}} = \frac{3 + 2\,(3.5) + 3\,(4) + 2\,(4.5) + 5}{9} = \frac{36}{9} = 4$$

Showing $\mu_{\overline{X}} = \mu = 4$

(v) $$\sigma_{\overline{X}} = \frac{(3-4)^2 + 2\,(3.5-4)^2 + 3\,(4-4)^2 + 2\,(4.5-4)^2 + (5-4)^2}{9}$$

$$= \frac{3}{9} = \frac{1}{3}$$

Therefore, $\sigma_{\overline{X}} = 0.5773$

Example 3.3 A random sample of size 2 is drawn from the population 3, 4, 5. Find (*i*) population mean (*ii*) Population S.D. (*iii*) Sampling distribution (SD) of means (*iv*) the mean of SD of means (*v*) S.D of SD means? Solve the problem without replacement?

[**Ans.** 0.4082]

Solution: (*i*) 4

(*ii*) $\sigma = 0.8164$

(*iii*) Sampling without replacement finite population thetoal number of samples without replacement is $N_{C_n} = {}^3C_2 = 3$ the three saples are (3, 4), (3, 5) (4, 5) and their means are 3.5, 4. 4.5

(*iv*) $\mu_{\overline{X}}$ = mean of smpling distribution of means $= \dfrac{(3.5)+4+4.5}{3} = \dfrac{12}{3} = 4 = \mu.$

(*v*) $$\sigma^2_{\overline{X}} = \frac{(3.5-4)^2+(4-4)^2+(4.5-4)^2}{3} = \frac{2.(0.5)^2}{3}$$

$$\sigma_{\overline{X}} = 0.4082.$$

Hence the solution.

Example 3.4 Find $P(\overline{X} > 66.75)$ if a random sample size 36 is drawn from an infinite population with mean $\mu = 63$ and SD $\delta = 9$. [**Ans.** 0.0062]

Solution: Let $$z = \frac{66.75-63}{9/\sqrt{36}} = 2.5$$

Hence, $P(\overline{X} > 66.75) = P(Z > 2.50) = 0.0062.$

Hence the solution.

Example 3.5 Determine the probability that mean breaking strength of cables produced by company 2 will be (i) at least 600N more than (ii) at least 450 N more than the cables produced by company 1, if 100 cables of brand 1 and 50 cables of brand 2 are tested.

Company	**Mean breaking strength**	**SD**	**Sample size**
1	4000 N	300 N	100
2	4500 N	200 N	50

[**Ans.** 0.8869]

Solution: $\mu_{X_2-X_1} = \mu_{\overline{X_2}-\overline{X_1}} = 4500 - 4000 = 500$ N

$$\sigma_{\overline{X_2}-\overline{X_1}} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(200)^2}{50}+\frac{(300)^2}{100}} = \sqrt{1700} = 41.23$$

(*i*) $P(\overline{X_2}-\overline{X_1}) > 600) = P\left(Z > \dfrac{600-500}{41.23}\right) = P(Z > 2.4254) = 0.0078$

(*ii*) $P(\overline{X_2}-\overline{X_1} > 450) = P\left(Z > \dfrac{450-500}{41.23}\right) = P(Z > -1.2127) = 0.8869.$

Hence the solution.

Example 3.6 Let $\overline{X_1}$ and $\overline{X_2}$ be the average drying time of two types of oil paints 1 and 2 for samples size $n_1 = n_2 = 18$. Suppose $\sigma_1 = \sigma_2 = 1$. Find the value of $P(\overline{X_1}-\overline{X_2} > 1)$, assuming that mean drying time is equal for the two types of oil paints. [**Ans.** 0.0013]

Solution: $\sigma^2_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{1}{18} + \frac{1}{18}} = \sqrt{\frac{1}{9}} = \frac{1}{3}$

$$P\left(X_1 - X_2\right) = P\left(Z > \frac{1 - (\mu_1 - \mu_2)}{\sigma\left(\overline{X}_1 - \overline{X}_2\right)}\right) = P\left(Z > \frac{1}{1/3}\right) = P(Z > 3) = 1 - 0.9987 = 0.0013$$

Hence, the solution.

Example 3.7 A company claims that the mean life time of tube lights is 500 hours. Is the claim of the company tenable if a random sample of 25 tube lights produced by the company has mean 518 hours and SD 40 hours. [**Ans.** 2.492]

Solution: Given $\overline{x} = 518$ hrs. $n = 25$, $s = 40$, $\mu = 500$

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{518 - 500}{40/\sqrt{25}} \text{ since, } t = 2.25 < t_{0.01, \upsilon = 24} = 2.492$$

Accept the claim of the company. Hence, the solution.

Example 3.8 Determine the probability that the variance of the first sample of size $n_1 = 9$ will be at least 4 times as large as the variance of the second sample of size $n_2 = 16$ if the two samples are independent random samples from a normal population.

[**Ans.** 0.01]

Solution: $F = \frac{s_1^2}{s_2^2} = 4 =$ and $\nu_1 = n_1 - 1 = 9 - 1 = 8$, $\nu_2 = n_2 - 1 = 16 - 1 = 15$

From table $F_{0.01}$ (8,16) = 4

Hence, the desired probability is 0.01

Hence the solution.

Example 3.9 Is there reason to believe that the life expected of Group A and Group B is same or not from the following data:

Group A	34	39.2	46.1	48.7	49.4	45.9	55.3	42.7	43.7	56.6
Group B	49.7	55.4	57.0	54.2	50.4	44.2	53.4	57.5	61.9	58.2

[**Ans.** 1.63]

Solution: Given data $s_A^2 = \frac{1}{8}\left(18527.78 - \frac{(405)^2}{9}\right) = 37.848$

$$s_B^2 = \frac{1}{10}\left(32799.918 - \frac{(598.5)^2}{11}\right) = 23.607$$

$$F = \frac{s_A^2}{s_B^2} = \frac{37.848}{23.607} = 1.63$$

clearly, variances empectancy is same for Group A and Group B. Hence, the solution.

Example 3.10 A random sample of size 25 from a normal population has the mean $\overline{x} = 47.5$ and the standard deviation $s = 8.4$. does this information tend to support of refute the claim that the mean of the population is $\mu = 42.1$? [**Ans.** $t = 3.21$]

Solution: Given $n = 25$, $\overline{x} = 47.5, \mu = 42.1$, $s = 8.4$ we have from t-distribution $t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{47.5 - 4.1}{8.4/\sqrt{25}}$ (or) $t = 3.21$. This value of t has 24 degrees of freedom. From the table of t-distribution for $\nu = 24$, we get probability that t will exceed 2.797 is 0.005. Then the probability of getting a value greater than 3.21 is negligible. Hence we conclude that the information given in the data of this example tend to refute the claim that the mean of the population is $\mu = 42.1$. Hence the solution.

Example 3.11 In 16 hour ten runs, the gasoline consumption of an engine averaged 16.4 gallons with a. s.d. of 2.1 gallons. Test the claim that the average gasoline consumption of this engine is 12.0 gallons per hour. [**Ans.** $t = 8.38$]

Solution: substituting $n = 16$, $\mu = 12.0$, $\overline{x} = 16.4$ and $s = 21$ into the formula for $t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{16.4 - 12.4}{2.1/\sqrt{16}}$, but from the table for $\nu = 15$ the probability of getting a value of t greater than 2.947 is 0.005. the probability of getting a value greater than 8 must be negligible. Thus, it would seem reasonable to conclude that the true average hourly gasoline consumption of the engine exceeds 12.0 gasoline. Hence, the solution.

Example 3.12 Suppose that the thickness of a part used in a semiconductor is its critical dimension, and that process of manufacturing these parts is considered to be under control if the true version among the thickness of the parts is given by a standard deviation not greater than $\sigma = 0.60$ thousandth of an inch. To keep a check on the process, random samples of size $n = 20$ are taken periodically, and is regarded to be "out of control" if the probability that s^2 will take on a value greater than or equal to the observed sample value is 0.01 or less even though $\sigma = 0.60$ what can one conclude about the process if the standard deviation of such a periodic random sample is $s = 0.84$ thousandth of an inch?

[**Ans.** 37.24]

Solution: The process will be declared 'out of control' if $\frac{(n-1)s^2}{\sigma^2}$ with $n = 20$ and $\sigma = 0.60$ exceeds $\chi^2_{0.01,19} = 36.91$, since $\frac{(n-1)s^2}{\sigma^2} = \frac{19(0.84)^2}{(0.60)^2} = 37.24$ exceeds 36.191, the process is declared out of control. Of course it is assumed here that the sample may be regarded as a random sample from a normal population. Hence the solution.

Example 3.13 A soft-drink vending machine is set so that the amount of drink dispensed is a random variable with a mean of 200 millilitres and a standard deviation of 15 millilitres'. What is the probability that the average (mean) amount dispensed in a random sample size of 36 at least 204 millilitres?

Solution: The distribution of $\bar{X}$ has the mean $\mu_{\bar{X}} = 200$ and the standard deviation $\sigma_{\bar{X}} = \frac{15}{\sqrt{36}} = 2.5$, and according to the central limit theorem, this distribution is approximately normal. And $Z = \frac{204 - 200}{2.5} = 1.6$.

Then $P(\bar{x} \geq 204) = P(Z \geq 1.6) = 0.5000 - 0.4452 = 0.0548$. Hence the solution.

Example 3.14 If two independent random sample of size $n_1 = 7$ and $n_2 = 13$ are taken from a normal population what is the probability that the variance of the first sample will be at least three times as large that of the second sample?

Solution: $F_{0.05}(v_1 = 6, v_2 = 12) = 3$ thus the desired probability is 0.05. Hence the solution.

Example 3.15: The claim that the variance of a normal population is $\sigma^2 = 21.3$ is rejected if the variance of a random sample of size 15 exceeds 39.74. What is the probability that the claim will be rejected even though $\sigma^2 = 21.3$? [**Ans.** 0.025]

Solution: $n = 15$, $\sigma^2 = 21.3$, $s^2 = 39.74$

$$\chi^2 = df \times \frac{s^2}{\sigma^2} = \frac{14(39.74)}{21.3} = 26.120$$

$$P(s^2 > 39.74) = P(\chi^2 > 26.120)$$

As we from χ^2 table

$$\chi^2_{0.025,14} = 26.119$$

$$P(s^2 > 39.74) = 0.025$$

Therefore, probability that the claim will be rejected is 0.0025. Hence, the solution.

Example 3.16 An electronic company manufactures resistors that have a mean resistance of 100 Ω and a standard deviation of 10 Ω. The distribution of resistance is normal. Find the probability that a random sample 25 resistors will have an average resistance less than 95 Ω? [**Ans.** 0.0062]

Solution: $n = 25$, $\mu = 100\ \Omega$, $\sigma = 10\ \Omega$ so $\mu_{\bar{X}} = 100$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{25}} = 2$

For = 95, $z = \frac{x - \mu}{\sigma} = \frac{95 - 100}{2} = -2.5$

Hence $P(\bar{X} < 95) = P(Z < -2.5) = F(-2.5) = 1 - F(2.5) = 1 - 0.9938 = 0.0062$

Hence, he solution.

Example 3.17 The mean voltage of a battery is 15 volt and s.d.is 0.2 volt. What is the probability that four such batteries connected in series will have a combined voltage of 60.8 or more volts? [**Ans.** 0.0228]

Solution: Let, the mean voltage of a batteries 1, 2, 3, 4 be $\overline{x_1}, \overline{x_2}, \overline{x_3}, \overline{x_4}$ the mean of the series of the four batteries connected is

$$\mu_{\overline{x_1}+\overline{x_2}+\overline{x_3}+\overline{x_4}} = \mu_{\overline{x_1}} + \mu_{\overline{x_2}} + \mu_{\overline{x_3}} + \mu_{\overline{x_4}} = 15 + 15 + 15 + 15 = 60$$

$$\sigma_{\overline{x_1}+\overline{x_2}+\overline{x_3}+\overline{x_4}} = \sqrt{\sigma^2_{\overline{x_1}} + \sigma^2_{\overline{x_2}} + \sigma^2_{\overline{x_3}} + \sigma^2_{\overline{x_4}}} = \sqrt{4(0.2)^2} = 0.4$$

Let X be the combined voltage of the series. When $X = 60.8, Z = \dfrac{X-\mu}{\sigma} = \dfrac{60.8-60}{0.4} = 2$

Then the probability that the combined voltage is more than 60.8 is given by $P(X \geq 60.8) = P(Z \geq 2) = 0.0228$. Hence the solution.

Example 3.18 Certain ball bearings have a mean weight of 5.02 ounces and standard deviation of 0.30 ounces. Find the probability that a random sample of 100 ball bearings will have a combined weight between 496 and 500 ounces? [**Ans.** 0.2318]

Solution: $\mu = 5.02, \sigma = 0.30, n = 100$

$$\mu_{\overline{X}} = \mu = 5.02,\ \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.30}{\sqrt{100}} = 0.03$$

$$\begin{aligned} P(4.96 < \overline{X} < 0.5) &= P\left(\frac{4.96-5.02}{0.03} < Z < \frac{5-5.02}{0.03}\right) = P(-2 < Z < -0.66) \\ &= F(-0.66) - F(-2) = F(2) - F(0.66) \\ &= 0.9772 - 0.7454 = 0.2318 \end{aligned}$$

Hence, the solution.

Example 3.19 A manufacturer of fuses claims that with a 20% overload, the fuses will blow in 12.40 minutes on the average. To test the claim, a sample of 20 of the fuses was subjected to a 20% overload, and the times it took them to blow had a mean of 10.63 minutes and a s.d. of 2.48 minutes. If it can be assumed that the data constitute a random sample from a normal population, do they tend to support or refute the manufacturer's claim? [**Ans.** –3.19]

Solution: $n = 20, \mu = 12.40, = 10.63, s = 2.48$ then $t = \dfrac{\overline{x}-\mu}{s/\sqrt{n}} = \dfrac{10.63-12.40}{2.48/\sqrt{20}}$

Date refutes the producer's claim since $t = -3.19 < -2.861$ with probability $\alpha = 0.005$. Hence, the solution.

Example 3.20 Show that for random samples of size n from a normal population with the variance σ^2, the sampling distribution of σ^2 has the mean σ^2 and the variance $\frac{2\sigma^4}{n-1}$?

Solution: We have $E\left(\frac{(n-1)s^2}{\sigma^2}\right) = n-1 \Rightarrow E(s^2) = \left(\frac{(n-1)\sigma^2}{n-1}\right) = \sigma^2$

$$Var\left(\frac{(n-1)s^2}{\sigma^2}\right) = 2(n-1)$$

$$Var(s^2) = \frac{\sigma^4 \times 2(n-1)}{(n-1)^2} = \frac{2\sigma^4}{(n-1)}$$

Hence, the solution.

Example 3.21 If $S_1{}^2$ and $S_2{}^2$ are the variances of independent random samples of size $n_1 = 10$ and $n_2 = 15$ from normal population with equal variances find $P(S_1{}^2/\,S_2{}^2 < 4.03)$?

[**Ans.** 0.99]

Solution: Let and $P = 1 - P(F > 4.03)$ with 9 and 14 d.o.f.

From table $F_{0.01}(9, 14) = 4.03$ then the probability $= 1 - 0.01 = 0.99$ Hence, the solution.

Example 3.22 A random sample of size $n = 25$ from a normal population has the mean $\overline{X} = 47$ and the standard deviation $\sigma = 7$. It we base our decision on the statistic, can we say that the given information supports the conjecture that the mean of the population is $\mu = 42$?

Solution: $f = \frac{47-42}{7/\sqrt{25}}$ since, 3.57 exceeds $t_{0.005,24} = 2.797$ for $\nu = 24$.

Clearly that the result is highly unlikely and conjecture is probably false.
Hence, the solution.

Example 3.23 The claim that the variance of a normal population is $\sigma^2 = 4$ is to be rejected if the variance of a random sample of size 9 exceeds 7.7535. What is the probability that this claim will be rejected even though $\sigma^2 = 4$?

[**Ans.** 0.5]

Solution: given $\sigma^2 = 4$, $n = 9$, $y = \frac{8s^2}{4} = 2\,s^2$

$P(y \geq 2\,(7.7535)) = P(y \geq 15.507)$ with 8 d.o.f. $= 0.5$ (χ^2-distribution table)

Hence, the solution.

Example 3.24 A random sample of size $n = 12$ from a normal population $\overline{x} = 27.8$ has the mean and the variance $\sigma^2 = 3.24$. If we base our decision on the t statistic can we say that the given information supports the claim that the mean of the population is $\mu = 28.5$?

[**Ans.** –1.347]

Solution: The statistic is $\frac{27.8-28.5}{1.8/\sqrt{12}} = -\frac{0.7}{\frac{1.8}{3.464}} = -1.347$ since this is fairly small and close to $-t_{0,10,11}$ the data tend to support the claim. Hence, the solution.

Example 3.25 The distribution of annual earnings of all bank letters with five years experience is skewed negatively. This distribution has a mean of ₹ 19000 and a standard deviation of ₹ 2000. If we draw a random sample of 30 tellers, what is the probability that the earnings will average more than ₹ 19750 annually? [**Ans.** 0.0202]

Solution: $\bar{X} = 19000$, $\mu = 30$, $\sigma = 2000$, standard error of the mean $(\sigma_x) = \sigma/\sqrt{n} = 2000/\sqrt{30}$ = consider the standard normal probability distribution, as follows: $Z = \frac{\bar{x}-\mu}{\sigma_x} = \frac{19750-19000}{365.16} = 2.05.$

Now P(earnings will average more than ₹ 19750 annually)

$$\begin{aligned} &= P(\bar{X} > 19750) \\ &= P(Z > 2.05) = 1 - P(Z \leq 2.05) \\ &= 1 - F(2.05) \\ &= 1 - 0.9798 = 0.0202 \end{aligned}$$

Therefore we have determined that there is slightly more than a 2% chance of average earnings more than ₹ 19750 annually in a group of 30 letters. Hence, the solution.

Example 3.26 If a gallon can of paint covers on the average 513.3 square feet ($Ft^{2.}$) with a standard deviation (s.d.) of 31.5 square feet ($Ft^{2.}$). what is the probability that the mean area covered by a sample of 40 of these 1 gallon cans will be anywhere from 510 to 520 square feet($Ft^{2.}$)? [**Ans.** 0.6553]

Solution: $n = 40$, $\mu = 513.3$ and $\sigma = 31.5$

Let $$Z = \frac{\bar{x}-\mu}{\sigma_x} = \frac{510-513.3}{31.5/\sqrt{40}} = -0.66$$

And $$Z = \frac{\bar{x}-\mu}{\sigma_x} = \frac{520-513.3}{31.5/\sqrt{40}} = 1.34$$

$$\begin{aligned} P(510 < \bar{X} < 520) &= P(-0.66 < Z < 1.34) \\ &= F(1.34) - F(-0.66) \\ &= F(1.34) - 1 + F(0.66) \\ &= 0.9099 - 1 + 0.7454 \\ &= 0.6553 \end{aligned}$$

We obtain the probability 0.6553 note that if turned out to be much less than 513.3, say less than 500 this might cause serious doubt whether the sample actually came from a population having $\mu = 513.3$ and $\sigma = 31.5$. the probability of obtaining such a small value *i.e.*, $Z < -2.67$ is only 0.0038. Hence, the solution.

Example 3.27 A random sample of 100 is taken from an infinite population having the mean $\mu = 76$ and the variance $= \sigma^2 = 256$. Find the probability that X will be between 75 and 78? [**Ans.** 0.6268]

Solution: $n = 100$, $\mu = 76$ and $\sigma = 256$

$$
\begin{aligned}
P(75 < \bar{X} < 78) &= P\left(\frac{75-76}{1.6} < Z < \frac{78-76}{1.6}\right) = P(-0.625 < Z < 1.25) \\
&= F(1.25) - F(-0.625) \\
&= F(1.25) - 1 + F(0.625) \\
&= 0.8944 - 1 + 0.7324 \\
&= 0.6268
\end{aligned}
$$

Hence, the solution.

Example 3.28 If two independent random samples of size $n_1 = 13$ and $n_2 = 7$ are taken from a normal population. What is the probability that the variance of the first sample will be atleast four times as that of the second sample? [**Ans.** 4.00]

Solution: Given $n_1 = 13$ and $\nu_1 = n_1 - 1 = 12$; $N_2 = 7$ and $\nu_2 = n_2 - 1 = 6$

$s_1{}^2 = 4s_2{}^2$ Now $F = \frac{s_1^2}{s_2^2} = \frac{4s_2^2}{s_2^2} = 4.0$. This value of F follows F-distribution with $\nu_1 = n_1 - 1 = 12$ and $\nu_2 = n_2 - 1 = 6$ degrees of freedom Hence from tables we get

$F_{0.05}(12,6) = 4.00$ Hence, the required probability is 0.05. Hence the solution.

Example 3.29 If two independent random samples of size $n_1 = 26$ and $n_2 = 8$ are taken from a normal population. What is the probability that the variance of the second sample will be atleast 2.4 times as that of the first sample? [**Ans.** 0.05]

Solution: Given $n_1 = 20$ and $n_2 = 8$, $\nu_1 = 19$ and $\nu_2 = 7$.

$$s_1{}^2 = (2.4)S_2{}^2$$

Now $$F_\alpha(19,7) = 2.4 \Rightarrow \alpha = 0.05$$

Hence, the solution.

Example 3.30 If the actual amount of instant coffee which a filing machine puts into '6-ounce' jars is r.v. having a normal distribution with SD 0.05 ounce and if only 3% of the jars are to contain less than 6 ounces of coffee, what must be the mean fill of these jars?

[**Ans.** $\mu = 6.094$]

Solution: Let X be the actual amount of coffee put into the jars, $X \sim N(\mu, 0.05)$

Given $P(X < 6) = 0.03$

$$P\left(-\infty < \frac{X-\mu}{0.05} < \frac{6-\mu}{0.05}\right) = 0.03$$

$$P(-\infty < Z \leq -z) = 0.5 - P(0 < Z \leq z)$$

$$0.5 - P\left(0 < Z < \frac{\mu-6}{0.05}\right) = 0.03$$

$$P\left(0 < Z < \frac{\mu-6}{0.05}\right) = 0.47 \text{ from table of areas } P(0 < Z < 1.808) = 0.47$$

Implies $\frac{\mu-6}{0.05} = 1.808 \Rightarrow \mu = 6.094$ ounces. Hence, the solution.

Example 3.31 A manufacturer of a certain type of synthetic fishing line has found from long experience of testing that the breaking strength of his product has an approximate normal distribution with a mean of 30 pounds(lbs.) and a standard deviation of 4 pounds (lbs.). A time and money saving change in the manufacturing process of the product is tried. A sample of 25 testing length pieces of the new process line is taken and tested with a resulting sample mean of 28 pounds(lbs.) What is the probability of obtaining a mean as low as 28 if the process has had no harmful effect on breaking strength? [**Ans.** 0.006]

Solution: Let X be the breaking strength of a randomly selected piece of line and if

$$X \sim N(30, 4) \text{ and } n = 25,\ \mu_{\bar{X}} \text{ (or } \bar{x}) = 30,\ \sigma_{\bar{X}} \text{ (or s)} = \frac{\sigma}{\sqrt{n}} = 0.8.$$

$$\text{Then } P(\bar{X} \leq 28) = P\left(\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{28-30}{0.8}\right) = P(Z \leq -2.5) = F(-2.5) = 1 - F(2.5)$$

$= 1 - 0.9938 = 0.006$. Thus there is a very small chance of obtaining a sample mean as low as 28. If there had been no change in the quality of the line due to the new process. Hence the solution.

Example 3.32 An Urn contains 1000 white and 2000 black balls. If X denotes the number of white balls when 300 balls are drawn without replacement, the find $P(180 < X < 120)$? [**Ans.** 0.9858]

Solution: Clearly $X \sim$ B.D^n = (300, 1/3)

If $p = P(\text{the ball drawn is white}) = 1/3$

$$\text{Mean } \mu = np = 300 \times 1/3 = 100$$

$$\text{Variance} = \sigma^2 = npq = 200/3$$

Since n = 30 is large the required probability is

$$P(80 < Z < 120) = P\left(\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{80-100}{200/\sqrt{3}} < Z < \frac{120-100}{200/\sqrt{3}}\right)$$

$$= P(-2.45 < Z < 2.45) = 0.9858$$

Hence the solution.

Example 3.33 Two movie theatres compete for 900 visitors. Suppose each visitor chooses one of the two balls independent of the choice of the other visitors; how many seats should each theatre have so that the probability of turning away any visitor for lack of seats is less than 1%? [**Ans.** 489]

Solution: Clearly $X \sim \text{B.D}^n = (900, 1/2)$

If
$$p = P(i = 1 \text{ to } 900) = 1/2$$

$$\text{Mean } \mu = np = 900 \times 1/2 = 450$$

$$\text{Variance} = \sigma^2 = npq = 900/2 \times 2 = 225$$

Since $n = 900$ is large the required probability is a $N(0, 1)$ random variable.

Now $P\left(-2.58 \leq \frac{X-450}{15} \leq 2.58\right) = 0.9902$

$P(411.3 < X < 488.7) = P(-2.45 < Z < 2.45) = 0.9858$. So the required number of seats is 489.

Hence, the solution.

Example 3.34 Let X be a random variable where μ_x is unknown as $\sigma_x^2 = 0.25$ *i.e.*, 1/4. Find out how large a random sample must be taken in order that the probability will be at test 0.95 and the sample mean $\overline{x}$ will lies within 0.25 of the population mean? [**Ans.** 80]

Solution: We have $\sigma_x^2 = 0.25$, $\varepsilon = 0.25$ and $1 - \frac{\sigma^2}{n\varepsilon^2} > 0.95$

Therefore $0.05 > \frac{\sigma^2}{n\varepsilon^2}$ and $n > \frac{\sigma^2}{0.05\varepsilon^2} = \frac{\sigma^2}{0.05\varepsilon^2} = \frac{1/4}{0.05 \times (0.25)^2} = 80.$

Hence, the solution.

Example 3.35 If a random sample of size n is selected from the finite population that consists of the integers 1, 2, 3, ..., N show that (*i*) the mean $\overline{X}$ is $\frac{N+1}{2}$ (*ii*) the variance σ^2 of $\overline{X}$ is $\frac{(N+1)(N-n)}{12n}$ (*iii*) the mean and the variance of $Y = n\overline{X}$ are $E(Y) = \frac{n(N+1)}{2}$ and the var $(Y) = \frac{n(N+1)(N-n)}{12}$.

Solution: (*i*) $$\mu = \overline{X} = \frac{1+2+3+\ldots+N}{N} = \frac{N(N+1)}{2N} = \frac{N+1}{2} \qquad \boxed{\therefore\ \mu = \frac{N+1}{2}}$$

(*ii*) $$\text{Var}\ (\overline{X}) = (\sigma^2) = \frac{1^2+2^2+3^2+\ldots+N^2}{N} - \frac{(N+1)(N+1)}{4}$$

$$= \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} \qquad \boxed{\therefore\ \sigma^2 = \frac{N^2-1}{12}}$$

$$\therefore \quad \boxed{\text{Var}(\overline{X}) = \frac{N^2-1}{12n} \cdot \frac{N-n}{N-1} = \frac{(N+1)(N-n)}{12}}$$

(iii) $$E(Y) = \mu_y = \frac{n(N+1)}{2}$$

$$\text{Var}\ (Y) = \frac{n^2(N+1)(N-n)}{12n} = \frac{n(N+1)(N+n)}{12}$$

$$\boxed{\therefore\ \text{Var}\ (Y) = \frac{n(N+1)(N+n)}{12}}$$

Example 3.36 How many different samples of size $n = 3$ can be drawn from a finite population of size (*a*) $N = 12$ (*b*) $N = 20$ (*c*) $N = 50$.[**Ans.** (*a*) 220, (*b*) 1140, (*c*) 19600]

Solution: (*a*) $^{12}C_3 = \frac{12.11.10}{3!} = 220;$

(*b*) $^{20}C_3 = \frac{20.19.18}{3!} = 1140;$

(*c*) $^{50}C_3 = \frac{50.49.48}{3!} = 19600;$

Hence, the solution.

Example 3.37 What is the probability of each possible sample if (*i*) a random sample of size $n = 4$ is to be drawn from a finite population of size $N = 12$ (*ii*) a random sample of size $n = 5$ is to be drawn from a finite population of size $N = 22$?

[**Ans.** (*a*) 1/495 (*b*) 1/77]

Solution: (*i*) $\frac{1}{N_{C_n}} = \frac{1}{12C_4} = \frac{1}{495}$ (*ii*) $\frac{1}{NC_n} = \frac{1}{12C_5} = \frac{1}{77}$

Hence, the solution.

Example 3.38 Independent random samples of size $n_1 = 30$ and $n_2 = 50$ are taken from two normal populations having the means $\mu_1 = 78$ and $\mu_2 = 78$ and the variances σ_1^2 and σ_2^2. Find the probability that the mean of the first sample will exceed that of the second sample by at least 4.8? [**Ans.** 0.2743]

Solution: Clearly $\mu_{\overline{x_1}-\overline{x_2}} = 78 - 75 = 3$

$$\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} = \sqrt{\frac{150}{30}+\frac{200}{50}} = 3$$

$$P((\overline{x_1}-\overline{x_2}) > 3) = P\left(Z > \frac{4.8-3.0}{3}\right) = P(Z > 0.6) = 0.2743.$$

Hence, the solution.

Example 3.39 If s_1 and s_2 are the variances of independent random samples of size $n_1 = 61$ and $n_2 = 31$ from normal population with $\sigma_1^2 = 12$ and $\sigma_2^2 = 18$. Find $P\left(\frac{s_1^2}{s_2^2} > 1.16\right)$. [**Ans.** 0.05]

Solution: Let $$F = \frac{s_1^2/12}{s_2^2/18} = 1.5\frac{s_1^2}{s_2^2}$$

Consider $$P\left(\frac{s_1^2}{s_2^2} > 1.16\right) = P\left(\frac{1.5\times s_1^2}{s_2^2} > 1.16\times 1.5\right)$$

$$= P(F > 1.74)$$

$F_{0.05}(60,30) = 1.74$ (from F-distribution Table). $P(F > 1.74) = 0.05$. Hence the solution.

PROBLEMS

1. Let X be a random variable where μ_x is unknown as $\sigma_x^2 = 0.20$ i.e., 1/5. Find out how large a random sample must be taken in order that the probability will be at test 0.99 and the sample mean $\overline{x}$ will lies within 0.20 of the population mean?
2. Standard error of the statistic sample mean $\overline{x}$ ____________.
3. If $x_1, x_2, x_3, \ldots,$ xn constitute a random sample from an infinite population with the mean μ and the variance σ^2 then $\mu(\overline{x}) =$ ____________ and $\sigma^2(\overline{x}) =$ ____________.
4. If $\overline{x}$ is the mean of a random sample from a finite population size N with the mean μ and the variance σ^2 then $\mu^2(\overline{x}) =$ ____________ and $\sigma^2(\overline{x}) =$ ______
5. Two movie theatres compete for 1000 visitors. If each visitor chooses one of the two balls independent of the choice of the other visitors; how many seats should each

theatre have so that the probability of turning away any visitor for lack of seats is less than 2%?

6. Independent random samples each with a $n_1 = 30$ and $n_2 = 50$ are drawn independently from two normal populations characterized by means $\mu_1 = 78$ and $\mu_2 = 78$ as well as variances σ_1^2 and σ_2^2. Determine the probability that the mean of the initial sample will exceed that of the second sample by a margin of at least 4.8?

7. What is the likelihood that the variance of a random sample, with a size of $n_1 = 26$, drawn from a normal population, will be at least 2.4 times greater than the variance of another independent random sample, with a size of $n_2 = 8$, taken from the same population?

8. What is the likelihood that, when drawing two independent random samples from a normal population with sizes of $n_1 = 20$ and $n_2 = 8$, the variance of the second sample will be at least 2.4 times greater than that of the first sample?

9. If s_1 and s_2 are the variances of independent random samples of size $n_1 = 61$ and $n_2 = 31$ from normal population with $\sigma_1^2 = 12$ and $\sigma^2_2 = 18$. Find $P\left(\frac{s_1^2}{s_2^2} > 1.16\right)$.

10. If the actual amount of instant coffee which a filing machine dispenses into '6-ounce' jars is r.v. having a normal distribution with SD 0.05 ounce. What should be the mean fill of the jars to ensure that only 3% of the jars contain less than 6 ounces of coffee?

11. An urn contains 1000 white and 2000 black balls. If X denotes the number of white balls when 300 balls are drawn without replacement, then find $P(180 < X < 120)$?

12. Two movie theatres compete for 900 visitors. Suppose each visitor chooses one of the two balls independent of the choice of the other visitors; how many seats should each theatre have so that the probability of turning away any visitor for lack of seats is less than 1%?

□□□

CHAPTER

4

CONFIDENCE INTERVALS

Chapter Outline

- Point Estimates • Confidence Intervals about the Mean (μ) when the Population Standard Deviation (σ) is Known • Confidence Intervals about the Mean (μ) when the Population Standard Deviation (σ) is Unknown • Confidence Intervals about the Population Proportion (p) • Confidence Interval Summary

In the preceding chapter we explord concept the that populations are characterized by descriptive measures called parameters. drawing inferences about these relies on sample statistics. We now want to estimate population parameters and assess the reliability of our estimates based on our knowledge of the sampling distributions of these statistics. A confidence interval is a statistical range or interval that provides a plausible range of values for an unknown population parameter, such as a population mean or proportion. It is constructed from sample data and is used to quantify the uncertainty or variability associated with estimating the true population parameter.

4.1 POINT ESTIMATES

Point estimate is a single value computed from the sample data that is used to estimate the population parameter of interest.

- The sample mean ($\overline{x}$) is a point estimate of the population mean (μ).
- The sample proportion $\hat{p}$ is the point estimate of the population proportion (p).

We use point estimates to construct confidence intervals for unknown parameters.

- A confidence interval is an interval of values instead of a single point estimate.
- The level of confidence corresponds to the expected proportion of intervals that will contain the parameter if many confidence intervals are constructed of the same sample size from the same population.
- Our uncertainty is about whether our particular confidence interval is one of those that truly contains the true value of the parameter.

Problem 4.1: We are 95% confident that our interval contains the population mean bear weight.

If we create 100 confidence intervals of the same size from the same population, we can expect 95 of them to contain the true parameters (the population mean weight). We also expect that five out of the all intervals would not contain the parameters.

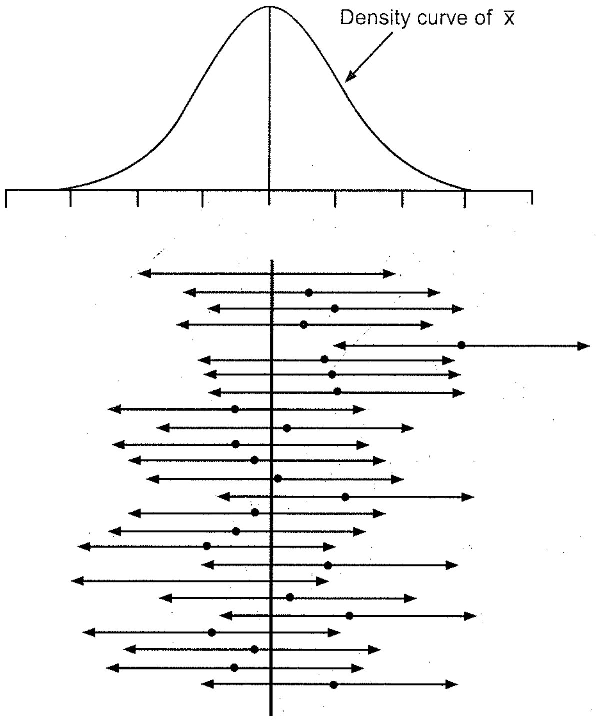


Fig. 4.1 Confidence intervals from twenty-five different samples

In this example, twenty-five samples from the same population gave these 95% confidence intervals. In the long term, 95% of all samples give an interval that contains μ, the true (but unknown) population mean.

Level of confidence is expressed as a percent.

- The compliment to the level of confidence is α (alpha), the level of significance.
- The level of confidence is described as $(1 - \alpha) \times 100\%$.

What does this really mean?

- We use a point estimate (*e.g.*, sample mean) to estimate the population mean.
- We attach a level of confidence to this interval to describe how certain we are that this interval actually contains the unknown population parameter.
- We want to estimate the population parameter, such as the mean (μ) or proportion (*p*).

$$\bar{x} - E < \mu < \bar{x} + E \quad \text{or} \quad \hat{p} - E < p < \hat{p} + E$$

where E is the margin of error.

The confidence is based on area under a normal curve. So the assumption of normality must be met

4.2 CONFIDENCE INTERVALS ABOUT THE MEAN (μ) WHEN THE POPULATION STANDARD DEVIATION (σ) IS KNOWN

A confidence interval takes the form of: point estimate ± margin of error.

The Point Estimate

- The point estimate comes from the sample data.
- To estimate the population mean (μ), use the sample mean ($\bar{x}$) as the point estimate.

The Margin of Error

- Depends on the level of confidence, the sample size and the population standard deviation.
- It is computed as $E = Z_{\alpha/2} \times \dfrac{\sigma}{\sqrt{n}}$ where $Z_{\alpha/2}$ is the critical value from the standard normal table associated with α (the level of significance).

The Critical Value $Z\alpha/_2$

- This is a Z-score that bounds the level of confidence.
- Confidence intervals are **always** two-sided and the Z-scores are the limits of the area associated with the level of confidence.

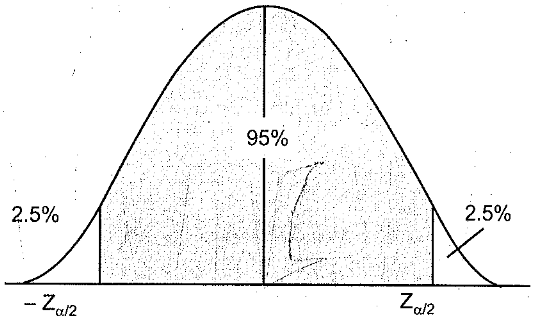


Fig. 4.2 The middle 95% area under a standard normal curve

- The level of significance (α) is divided into halves because we are looking at the middle 95% of the area under the curve.
- Go to your standard normal table and find the area of 0.025 in the body of values.
- What is the Z-score for that area?
- The Z-scores of ± 1.96 are the critical Z-scores for a 95% confidence interval.

Table 4.1: Common critical values (Z-scores)

Confidence Level	**α (level of significance)**	$Z\alpha/_2$
99%	1%	2.575
95%	5%	1.96
90%	10%	1.645

Construction of a confidence interval about μ when σ is known:

1. $Z_{\alpha/2}$ (critical value)
2. $E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ (margin of error)
3. $\bar{x} \pm E$ (point estimate ± margin of error)

Problem 4.2: Construct a confidence interval about this population mean.

Researchers have been studying *p*-loading in Jones Lake for many years. It is known that mean water clarity (using a Secchi disk) is normally distributed with a population standard deviation of $\sigma = 15.4$ in. A random sample of 22 measurements was taken at various points on the lake with a sample mean of $\bar{x} = 57.8$ in. The researchers want you to construct a 95% confidence interval for μ, the mean water clarity.

1. $Z_{\alpha/2} = 1.96$
2. $E = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{15.4}{\sqrt{22}} = 6.435$
3. $\bar{x} \pm E = 57.8 \pm 6.435$

95% confidence interval for the mean water clarity is (51.36, 64.24).

We can be 95% confident that this interval contains the population mean water clarity for Jones Lake.

Now construct a 99% confidence interval for μ, the mean water clarity, and interpret.

1. $Z_{\alpha/2} = 2.575$
2. $E = Z_{\alpha/2} \times \frac{15.4}{\sqrt{22}}$
3. $\bar{x} \pm E = 57.8 \pm 8.454$

99% confidence interval for the mean water clarity is (49.35, 66.25).

We can be 99% confident that this interval contains the population mean water clarity for Jones Lake.

As the level of confidence increased from 95% to 99%, the width of the interval increased. As the probability (area under the normal curve) increased, the critical value increased resulting in a wider interval.

4.3 CONFIDENCE INTERVALS ABOUT THE MEAN (μ) WHEN THE POPULATION STANDARD DEVIATION (σ) IS UNKNOWN

Typically, in real life we often don't know the population standard deviation (σ). We can use the sample standard deviation (s) in place of σ. However, because of this change, we can't use the standard normal distribution to find the critical values necessary for constructing a confidence interval.

The Student's *t*-distribution was created for situations when σ was unknown. Gosset worked as a quality control engineer for Guinness Brewery in Dublin. He found errors in his testing and he knew it was due to the use of *s* instead of σ. He created this distribution to deal with the problem of an unknown population standard deviation and small sample sizes. A portion of the *t*-table is shown below.

Table 4.2: Portion of the student's *t*-table.

Area in Right Tail

df	0.10	0.05	0.025	0.02	0.01	0.005
1	3.078	6.314	12.706	15.894	31.821	63.657
2	1.886	2.920	4.303	4.849	6.965	9.925
3	1.638	2.353	3.182	3.482	4.541	5.841
4	1.533	2.132	2.776	2.999	3.747	4.604
5	1.476	2.015	2.571	2.757	3.365	4.032

Problem 4.3 Find the critical value $t_{\alpha/2}$ for a 95% confidence interval with a sample size of $n = 13$.

Degrees of freedom (down the left-hand column) is equal to $n - 1 = 12$

$\alpha = 0.05$ and $\alpha/2 = 0.025$

Go down the 0.025 column to 12 *df*

$t_{\alpha/2} = 2.179$

The critical values from the students' *t*-distribution approach the critical values from the standard normal distribution as the sample size (n) increases.

Table 4.3: Critical values from the student's *t*-table.

n	Degree of freedom	$t_{0.25}$
11	10	2.228
51	50	2.009
101	100	1.984
1001	1000	1.962

Using the standard normal curve, the critical value for a 95% confidence interval is 1.96. You can see how different samples sizes will change the critical value and thus the confidence interval, especially when the sample size is small.

Construction of a Confidence Interval about μ when σ is Unknown

1. $t_{\alpha/2}$ critical value with $n-1$ *df*
2. $E = t_{\alpha/2} \times \dfrac{s}{\sqrt{n}}$
3. $\bar{x} \pm E$

Problem 4.4 Researchers studying the effects of acid rain in the Adirondack Mountains collected water samples from 22 lakes. They measured the pH (acidity) of the water and want to construct a 99% confidence interval about the mean lake pH for this region. The sample mean is 6.4438 with a sample standard deviation of 0.7120. They do not know anything about the distribution of the pH of this population, and the sample is small ($n < 30$), so they look at a normal probability plot.

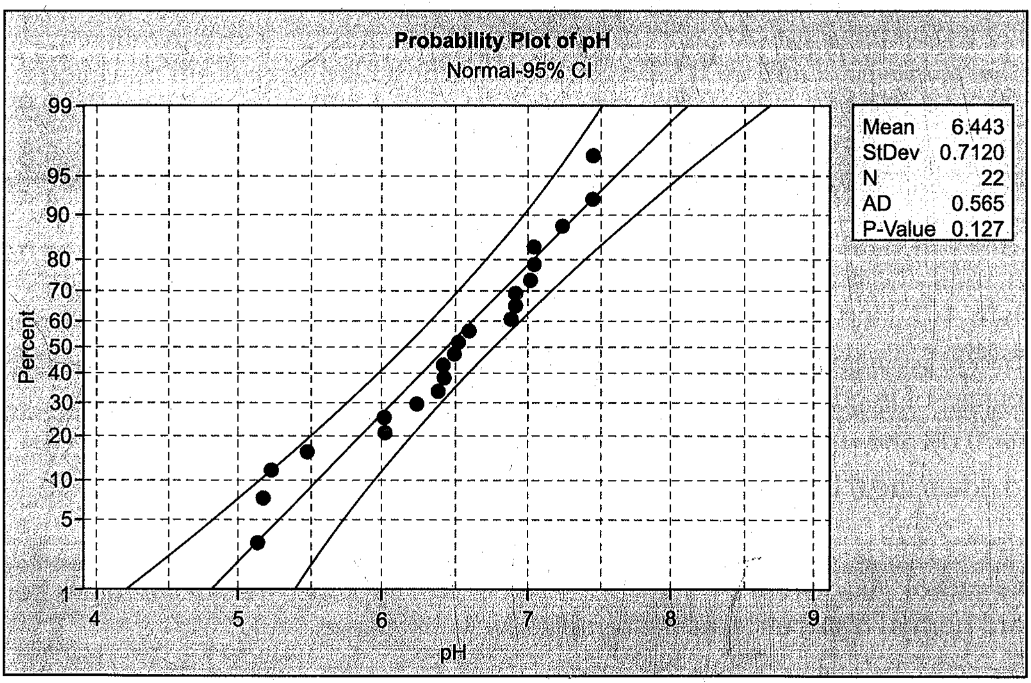


Fig. 4.3 Normal probability plot

The data is normally distributed. Now construct the 99% confidence interval about the mean pH.

1. $t_{\alpha/2} = 2.831$
2. $E = t_{\alpha/2} \times \dfrac{s}{\sqrt{n}} = = 2.831 \times \dfrac{0.7120}{\sqrt{22}}\ 0.4297$
3. $\bar{x} \pm E = 6.443 \pm 0.4297$

The 99% confidence interval about the mean pH is (6.013, 6.863).

We are 99% confident that this interval contains the mean lake pH for this lake population.

Now construct a 90% confidence interval about the mean pH for these lakes.

1. $t_{\alpha/2} = 1.721$
2. $E = t_{\alpha/2} \times \frac{s}{\sqrt{n}} = 21/721 \times \frac{0.7120}{\sqrt{22}} = 0.2612$
3. $\bar{x} \pm E = 6.443 \pm 0.2612$

The 90% confidence interval about the mean pH is (6.182, 6.704).

We are 90% confident that this interval contains the mean lake pH for this lake population.

Notice how the width of the interval decreased as the level of confidence decreased from 99 to 90%.

Construct a 90% confidence interval about the mean lake pH using Excel and Minitab.

4.4 CONFIDENCE INTERVALS ABOUT THE POPULATION PROPORTION (P)

Frequently, we are interested in estimating the population proportion , instead of the population mean. For example, you may need to estimate the proportion of trees infected with beech bark disease, or the proportion of people who support 'green' products. The parameter p can be estimated in the same ways as we estimated the population mean.

The Sample Proportion

The sample proportion is the best point estimate for the true population proportion. Sample proportion where x is the number of elements in the sample with the characteristic you are interested in, and n is the sample size.

The Assumption of Normality when Estimating Proportions

The assumption of a normally distributed population is still important, even though the parameter has changed.

Normality can be verified if:

$$n \times \hat{p} \times (1 - \hat{p}) \geq 10$$

Constructing a Confidence Interval about the Population Proportion

Constructing a confidence interval about the proportion follows the same three steps we have used in previous examples.

1. $Z_{\alpha/2}$ (critical value from the standard normal table)

2. $E = Z_{\alpha/2} \times \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}}$ (margin of error)
3. $\hat{p} \pm E$ (point estimate ± margin of error)

Problem 4.5 A botanist has produced a new variety of hybrid soybean that is better able to withstand drought. She wants to construct a 95% confidence interval about the germination rate (percent germination). She randomly selected 500 seeds and found that 421 have germinated.

First, compute the point estimate.

$$\hat{p} = \frac{x}{n} = \frac{421}{500}$$

Check normality:

$n \times \hat{p} \times (1 - \hat{p}) \geq 10 = 500 \times 0.842 \times (1 - 0.842) = 66.5 \geq 10.$

You can assume a normal distribution.

Now construct the confidence interval:

1. $Z_{\alpha/2} = 1.96$
2. $E = Z_{\alpha/2} \times \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}} = 1.96 \times \sqrt{\frac{0.842 \times (1-0.842)}{500}} = 0.032$
3. $\hat{p} \pm E = 0.842 \pm 0.032$

The 95% confidence interval for the germination rate is (81.0%, 87.4%).

We can be 95% confident that this interval contains the true germination rate for this population.

4.5 CONFIDENCE INTERVAL SUMMARY

Which method do I use?

The first question to ask yourself is: Which parameter are you trying to estimate? If it is the mean (μ), then ask yourself: Is the population standard deviation (σ) known? If yes, then follow the next 3 steps:

Confidence Interval about the Population Mean (μ) when σ is known

1. $Z_{\alpha/2}$ (critical value)
2. $E = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$ (margin of error)
3. (point estimate ± margin of error)

If no, follow these 3 steps:

Confidence Interval about the Population Mean (μ) when σ is Unknown

1. $t_{\alpha/2}$ critical value with $n - 1$ df from the student t-distribution
2. $E = t_{\alpha/2} \times \dfrac{\sigma}{\sqrt{n}}$
3. $\overline{x} \pm E$

If you want to construct a confidence interval about the population proportion, follow these 3 steps:

Confidence Interval about the Proportion

1. $Z_{\alpha/2}$ (critical value from the standard normal table)
2. $E = Z_{\alpha/2} \times \sqrt{\dfrac{\hat{p} \times (1-\hat{p})}{n}}$ (margin of error)
3. $\hat{p} \pm E$ (point estimate ± margin of error)

Remember that the assumption of normality must be verified.

PROBLEMS

1. When to use t distribution and when to use z distribution?

2. Construct a 95% confidence interval for the population standard deviation of a random sample of 15 men who have a mean weight of 165.2 pounds with a standard deviation of 13.5 pounds. Assume the population is normally distributed.

3. A random sample of 28 items is drawn from a population whose standard deviation is unknown. The sample mean is $\overline{x} = 900$ and the sample standard deviation is $s = 10$.

(*a*) Construct an interval estimate of μ with 98% confidence.

(*b*) Construct an interval estimate of μ with 98% confidence assuming that $s = 20$.

(*c*) Construct an interval estimate of μ with 98% confidence assuming that $s = 40$.

4. Determine the point estimate of the population proportion, the margin of error for the following confidence interval, and the number of individuals in the sample with the specified characteristic, x for the sample size provided. Lower bound = 0.253, upper bound = 0.797, $n = 1000$.

5. In a survey, the planning value for the population proportion is $p^* = 0.35$. How large a sample should be taken to provide a 95% confidence interval with a margin of error of 0.05?

6. A study showed that 0.2 of the population was divorced in Town A. How big a sample is required if we were to construct a 90% confidence interval with no more than 0.01 margin of error?

7. Suppose a random sample of size $n = 36$ has been selected from a population with $\sigma = 30$ and a sample mean of $\overline{x} = 205$.

   What is the 90% confidence interval estimate of μ?

8. In a continuing study of the amount MBA students are spending each term on textbooks, data were collected on $n = 81$ students. In previous studies, the population standard deviation has been $\sigma = \$24$.
   - (*a*) What is the margin of error at the 99% confidence level?
   - (*b*) If the mean from the most recent sample was $\overline{x} = \$288$, what is the 99% confidence interval estimate of the population mean μ?

9. Find the following probabilities using the t distribution with $n = 23$.
   - (*a*) What is the probability that t will be greater than or equal to 0? That is, what is $p(t \geq 0, df = 23 - 1 = 22)$?
   - (*b*) What is the probability that t will be greater than or equal to 2.074? That is, what is $p(t \geq 2.074, df = 23 - 1 = 22)$?
   - (*c*) What is the probability that t will be less than or equal to 2.819? That is, what is $p(t \leq 2.819, df = 23 - 1 = 22)$?
   - (*d*) What is the probability that t will be less than or equal to -1.321? That is, what is $p(t \leq 1.321, df = 23 - 1 = 22)$?
   - (*e*) What is the probability that t falls between -0.858 and 1.717? That is, what is the $p(-0.858 \leq t \leq 1.717)$?

❑❑❑

CHAPTER

5

HYPOTHESIS TESTING

Chapter Outline

• The Fundamentals of Hypothesis Testing • Hypothesis Test about the Population Mean (μ) when the Population Standard Deviation (σ) is Known • Hypothesis Test about the Population Mean (μ) when the Population Standard Deviation (σ) is Unknown • Hypothesis Test for a Population Proportion (p) • Hypothesis Test about a Variance • Inferences about the Differences of Two Populations • Inferences about Two Population Proportions • F-Test for Comparing Two Population Variances

The previous chapters introduced methods for organizing and summarizing sample data, and using sample statistics to estimate population parameters. This chapter introduces the next major topic of inferential statistics: hypothesis testing.

A hypothesis is a statement or claim about a property of a population.

5.1 THE FUNDAMENTALS OF HYPOTHESIS TESTING

When conducting scientific research, typically there is some known information, perhaps from some past work or from a long accepted idea. We want to test whether this claim is believable. This is the basic idea behind a hypothesis test:

- State what we think is true.
- Quantify how confident we are about our claim.
- Use sample statistics to make inferences about population parameters.

For example, past researches done tells us that the average lifespan for a humming bird is about four years. You have been studying the hummingbirds in the southeastern United States and find a sample mean lifespan of 4.8 years. Should you reject the known or accepted information in favor of your results? How confident are you in your estimate? At what point would you say that there is enough evidence to reject the known information and support your alternative claim? How far from the known mean of four years can the sample mean be before we reject the idea that the average lifespan of a hummingbird is four years?

Hypothesis testing is a procedure, based on sample evidence and probability, used to test claims regarding a characteristic of a population.

A hypothesis is a claim or statement about a characteristic of a population of interest to us. A hypothesis test is a way for us to use our sample statistics to test a specific claim.

Examples: The population mean weight is known to be 157 lb. We want to test the claim that the mean weight has increased.

Two years ago, the proportion of infected plants was 37%. We believe that a treatment has helped, and we want to test the claim that there has been a reduction in the proportion of infected plants.

5.1.1 Components of a Formal Hypothesis Test

The null hypothesis is a statement about the value of a population parameter, such as the population mean (μ) or the population proportion (p). It contains the condition of equality and is denoted as H_0 (H-naught).

$$H_0: \mu = 157 \text{ or } H_1: p \neq 0.37$$

The alternative hypothesis is the claim to be tested, the opposite of the null hypothesis. It contains the value of the parameter that we consider plausible and is denoted as H_1.

$$H_1: \mu > 157 \text{ or } H_1: p \neq 0.37$$

The test statistic is a value computed from the sample data that is used in making a decision about the rejection of the null hypothesis. The test statistic converts the sample mean ($\bar{x}$) or sample proportion ($\hat{p}$) to a Z- or t-score under the assumption that the null hypothesis is true. It is used to decide whether the difference between the sample statistic and the hypothesized claim is significant.

The *p*-value is the area under the curve to the left or right of the test statistic. It is compared to the level of significance (α).

The critical value is the value that defines the rejection zone (the test statistic values that would lead to rejection of the null hypothesis). It is defined by the level of significance.

The level of significance (α) is the probability that the test statistic will fall into the critical region when the null hypothesis is true. This level is set by the researcher.

The conclusion is the final decision of the hypothesis test. The conclusion must always be clearly stated, communicating the decision based on the components of the test. It is important to realize that we never prove or accept the null hypothesis. We are merely saying that the sample evidence is not strong enough to warrant the rejection of the null hypothesis. The conclusion is made up of two parts:

(1) Reject or fail to reject the null hypothesis, and (2) there is or is not enough evidence to support the alternative claim.

Option (1) Reject the null hypothesis (H_0). This means that you have enough statistical evidence to support the alternative claim (H_1).

Option (2) Fail to reject the null hypothesis (H_0). This means that you do NOT have enough evidence to support the alternative claim (H_1).

Similar to the US justice system, visualizing hypothesis testing offers a compelling comparision. In this comparison, the null hypothesis equates to the presumption of innocence for a defendant. The prosecuting attorney endeavors to establish the guilt of the defendant, corresponding to the alternative hypothesis. The jury can arrive at two possible conslusions: first, finding the defendant guilty (resulting in the rejection of the null hypothesis), and second, declaring the defendant not guilty (resulting in the failure to reject the null hypothesis). Importantly, it is crucial to recognize that 'not guilty' does not equate to a declaration of innocence. In the null hypothesis (innocence) in favor of the alternative hypothesis (guilt). In the latter scenario, the prosecutor lacked adequate avidence to reject the null hypothesis (innocence) and substantiate the alternative claim of guilt.

5.1.2 The Null and Alternative Hypotheses

There are three different pairs of null and alternative hypotheses:

Table 5.1: Different Type of Hypotheses

Two sided	Left-sided	Right-sided
$H_0: \mu = c$ $H_1: \mu \neq c$	$H_0: \mu = c$ $H_1: \mu < c$	$H_0: \mu = c$ $H_1: \mu > c$

where *c* is some known value.

5.1.3 A Two-sided Test

This tests whether the population parameter is equal to, versus not equal to, some specific value.

$$H_0: \mu = 12 \text{ vs. } H_1: \mu \neq 12$$

The critical region is divided equally into the two tails and the critical values are ± values that define the rejection zones.

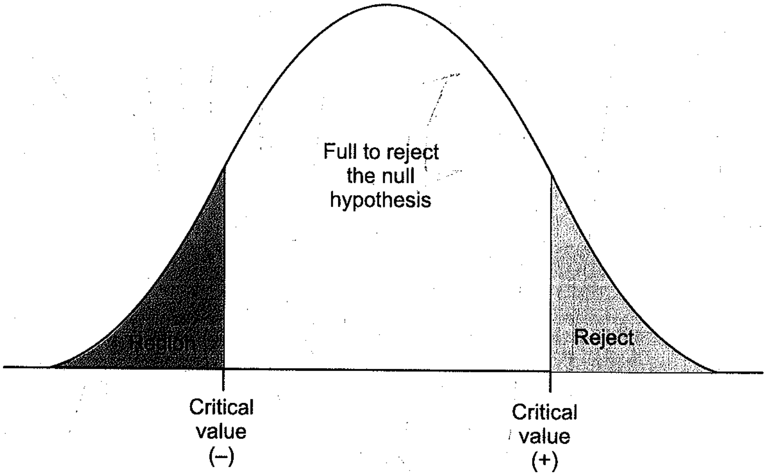


Fig. 5.1 The rejection zone for a two-sided hypothesis test

Example 5.1. A forester studying diameter growth of red pine believes that the mean diameter growth will be different if a fertilization treatment is applied to the stand.

$$H_0: \mu = 1.2 \text{ in./year}$$
$$H_1: \mu \neq 1.2 \text{ in./year}$$

This is a two-sided question, as the forester doesn't state whether population mean diameter growth will increase or decrease.

5.1.4 A Right-sided Test

This tests whether the population parameter is equal to, versus greater than, some specific value.

$$H_0: \mu = 12 \text{ vs. } H_1: \mu > 12$$

The critical region is in the right tail and the critical value is a positive value that defines the rejection zone.

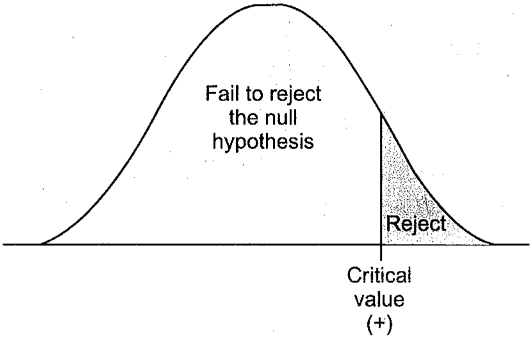


Fig. 5.2 The rejection zone for a right-sided hypothesis test

Example 5.2 A biologist believes that there has been an increase in the mean number of lakes infected with milfoil, an invasive species, since the last study five years ago.

$$H_0: \mu = 15 \text{ lakes}$$
$$H_1: \mu > 15 \text{ lakes}$$

This is a right-sided question, as the biologist believes that there has been an increase in population mean number of infected lakes.

5.1.5 A Left-sided Test

This tests whether the population parameter is equal to, versus less than, some specific value.

$$H_0: \mu = 12 \text{ vs. } H_1: \mu < 12$$

The critical region is in the left tail and the critical value is a negative value that defines the rejection zone.

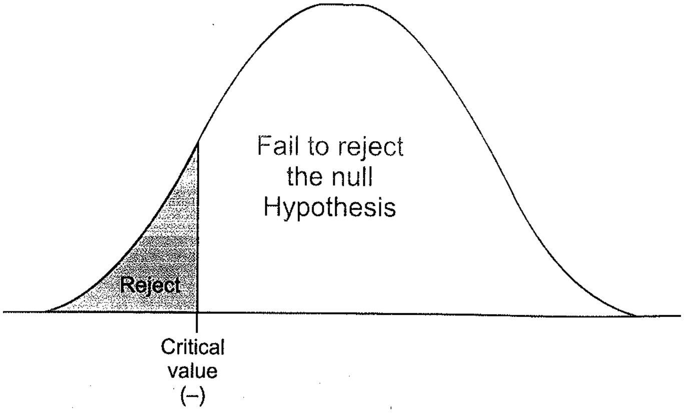


Fig. 5.3 The rejection zone for a left-sided hypothesis test

Example 5.3 A scientist's research indicates that there has been a change in the proportion of people who support certain environmental policies. He wants to test the claim that there has been a reduction in the proportion of people who support these policies.

This is a left-sided question, as the scientist believes that there has been a reduction in the true population proportion.

5.1.6 Statistically Significant

When the observed results (the sample statistics) are unlikely (a low probability) under the assumption that the null hypothesis is true, we say that the result is statistically significant, and we reject the null hypothesis. This result depends on the level of significance, the sample statistic, sample size, and whether it is a one- or two-sided alternative hypothesis.

5.1.7 Types of Errors

When testing, we arrive at a conclusion of rejecting the null hypothesis or failing to reject the null hypothesis. Such conclusions are sometimes correct and sometimes incorrect (even when we have followed all the correct procedures). We use incomplete sample data to reach a conclusion and there is always the possibility of reaching the wrong conclusion. There are four possible conclusions to reach from hypothesis testing. Of the four possible outcomes, two are correct and two are NOT correct.

Reality

Table 5.2 Possible outcomes from a hypothesis test

	H_0 is True	H_1 is True
Do Not Reject H_0	Correct Conclusion	Type II Error
Reject H_0	Type I Error	Correct Conclusion

A Type I error is when we reject the null hypothesis when it is true. The symbol α (alpha) is used to represent Type I errors. This is the same alpha we use as the level of significance. By setting alpha as low as reasonably possible, we try to control the Type I error through the level of significance.

A Type II error is when we fail to reject the null hypothesis when it is false. The symbol β (beta) is used to represent Type II errors.

In general, Type I errors are considered more serious. One step in the hypothesis test procedure involves selecting the significance level (α), which is the probability of rejecting the null hypothesis when it is correct. So the researcher can select the level of significance that minimizes Type I errors. However, there is a mathematical relationship between α, β, and n (sample size).

As α increases, β decreases

As α decreases, β increases

As sample size increases (n), both α and β decrease.

The natural inclination is to select the smallest possible value for α, thinking to minimize the possibility of causing a Type I error. Unfortunately, this forces an increase in Type II errors. By making the rejection zone too small, you may fail to reject the null hypothesis, when, in fact, it is false. Typically, we select the best sample size and level of significance, automatically setting β.

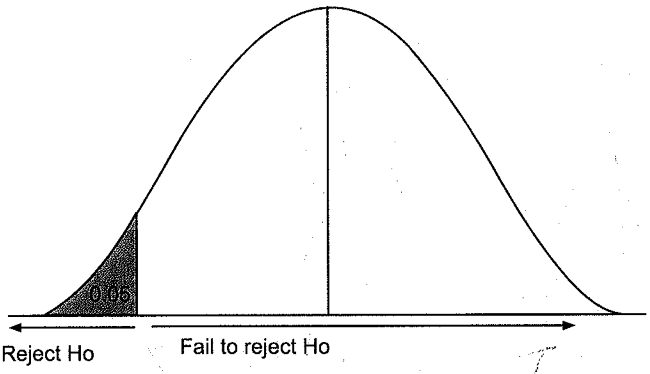


Fig. 5.4. Type 1 error.

5.1.8 Power of the Test

A Type II error (β) is the probability of failing to reject a false null hypothesis. It follows that 1-β is the probability of rejecting a false null hypothesis. This probability is identified as the **power** of the test, and is often used to gauge the test's effectiveness in recognizing that a null hypothesis is false.

The probability that at a fixed level α significance test will reject H_0, when a particular alternative value of the parameter is true is called the power of the test.

Power is also directly linked to sample size. For example, suppose the null hypothesis is that the mean fish weight is 8.7 lb. Given sample data, a level of significance of 5%, and an alternative weight of 9.2 lb., we can compute the power of the test to reject $\mu \doteq 8.7$ lb. If we have a small sample size, the power will be low. However, increasing the sample size will increase the power of the test. Increasing the level of significance will also increase power. A 5% test of significance will have a greater chance of rejecting the null hypothesis than a 1% test because the strength of evidence required for the rejection is less. Decreasing the standard deviation has the same effect as increasing the sample size: there is more information about μ.

5.2 HYPOTHESIS TEST ABOUT THE POPULATION MEAN (μ) WHEN THE POPULATION STANDARD DEVIATION (σ) IS KNOWN

We are going to examine two equivalent ways to perform a hypothesis test: the classical approach and the *p*-value approach. The **classical approach** is based on standard deviations. This method compares the test statistic (*Z*-score) to a critical value (*Z*-score) from the standard normal table. If the test statistic falls in the rejection zone, you reject the null hypothesis. The **p-value approach** is based on area under the normal curve. This method compares the area associated with the test statistic to alpha (α), the level of significance (which is also area under the normal curve). If the *p*-value is less than alpha, you would reject the null hypothesis.

Both methods must have:

- Data from a random sample.
- Verification of the assumption of normality.
- A null and alternative hypothesis.
- A criterion that determines if we reject or fail to reject the null hypothesis.
- A conclusion that answers the question.
- There are four steps required for a hypothesis test:
- State the null and alternative hypotheses.
- State the level of significance and the critical value.
- Compute the test statistic.
- State a conclusion.

5.2.1 Testing a Hypothesis using classical Approach

The classical approach is based on standard deviations. This method compares the test statistic (Z-score) to a critical value (Z-score) from the standard normal table. If the test statistic falls in the rejection zone, you reject the null hypothesis

Problem 5.1 *A Two-sided Test*

A forester studying diameter growth of red pine believes that the mean diameter growth will be different from the known mean growth of 1.35 inches/year if a fertilization treatment is applied to the stand. He conducts his experiment, collects data from a sample of 32 plots, and gets a sample mean diameter growth of 1.6 in./year. The population standard deviation for this stand is known to be 0.46 in./year. Does he have enough evidence to support his claim?

Step (1) State the null and alternative hypotheses.

$$H_0: \mu = 1.35 \text{ in./year}$$
$$H_1: \mu \neq 1.35 \text{ in./year}$$

Step (2) State the level of significance and the critical value.

We will choose a level of significance of 5% ($\alpha = 0.05$).

For a two-sided question, we need a two-sided critical value $-Z_{\alpha/2}$ and $+Z_{\alpha/2}$.

The level of significance is divided by 2 (since we are only testing "not equal"). We must have two rejection zones that can deal with either a greater than or less than outcome (to the right (+) or to the left (–)).

We need to find the *Z*-score associated with the area of 0.025. The red areas are equal to $\alpha/2 = 0.05/2 = 0.025$ or 2.5% of the area under the normal curve.

Go into the body of values and find the negative *Z*-score associated with the area 0.025.

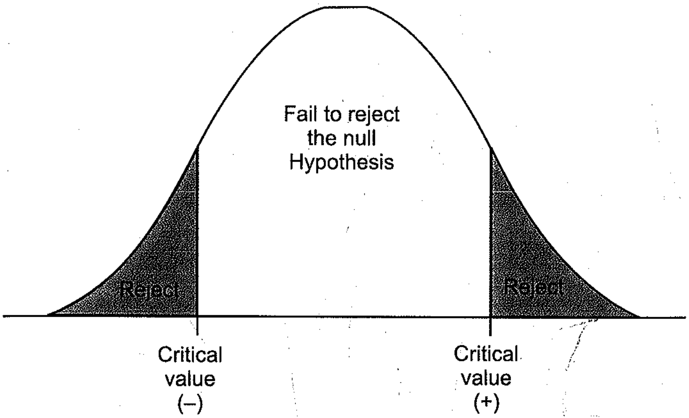


Fig. 5.5. The rejection zone for a two-sided test.

The negative critical value is –1.96. Since the curve is symmetric, we know that the positive critical value is 1.96.

± 1.96 are the critical values. These values set up the rejection zone. If the test statistic falls within these shaded rejection zones, we reject the null hypothesis.

Step (3) Compute the test statistic.

The test statistic is the number of standard deviations the sample mean is from the known mean. It is also a *Z*-score, just like the critical value.

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

For this problem, the test statistic is

$$Z = \frac{1.6 - 1.35}{0.46 / \sqrt{32}}$$

Step (4) State a conclusion.

Compare the test statistic to the critical value. If the test statistic falls into the rejection zones, reject the null hypothesis. In other words, if the test statistic is greater than +1.96 or less than –1.96, reject the null hypothesis.

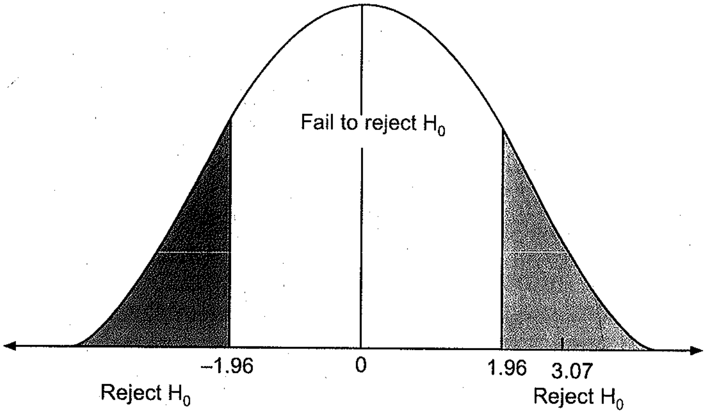


Fig. 5.6 The critical values for a two-sided test when $\alpha = 0.05$

In this problem, the test statistic falls in the shaded rejection zone. The test statistic of 3.07 is greater than the critical value of 1.96.We will reject the null hypothesis. We have enough evidence to support the claim that the mean diameter growth is different from (not equal to) 1.35 in./year.

Problem 5.2 *A Right-sided Test*

A researcher believes that there has been an increase in the average farm size in his state since the last study five years ago. The previous study reported a mean size of 450 acres with a population standard deviation (α) of 167 acres. He samples 45 farms and gets a sample mean of 485.8 acres. Is there enough information to support his claim?

Step (1) State the null and alternative hypotheses.

$$H_{0:}\ \mu = 450 \text{ acres}$$
$$H_{1:}\ \mu > 450 \text{ acres}$$

Step (2) State the level of significance and the critical value.

We will choose a level of significance of 5% ($\alpha = 0.05$).

For a one-sided question, we need a one-sided positive critical value $Z\alpha$.

The level of significance is all in the right side (the rejection zone is just on the right side).

We need to find the Z-score associated with the 5% area in the right tail.

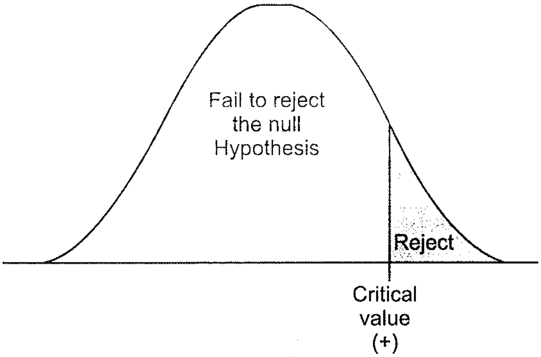


Fig. 5.7 Rejection zone for a right-sided hypothesis test

Go into the body of values in the standard normal table and find the Z-score that separates the lower 95% from the upper 5%.

The critical value is 1.645. This value sets up the rejection zone.

Step (3) Compute the test statistic.

The test statistic is the number of standard deviations the sample mean is from the known mean. It is also a Z-score, just like the critical value.

$$Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

For this problem, the test statistic is

$$Z = \frac{485.8 - 450}{167 / \sqrt{45}}$$

Step (4) State a conclusion.

Compare the test statistic to the critical value.

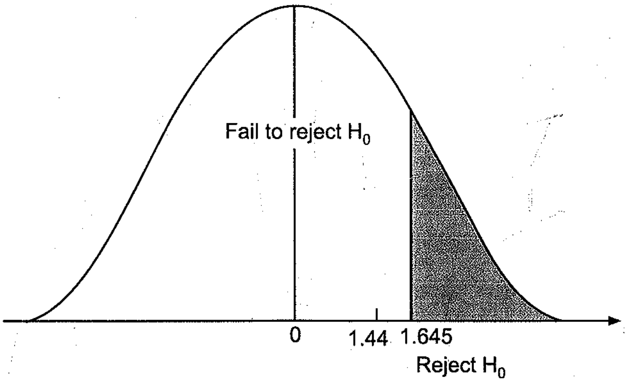


Figure 5.8 The critical value for a right-sided test when $\alpha = 0.05$

The test statistic does not fall in the rejection zone. It is less than the critical value.

We fail to reject the null hypothesis. We do not have enough evidence to support the claim that the mean farm size has increased from 450 acres.

Problem 5.3 *A Left-sided Test*

A researcher believes that there has been a reduction in the mean number of hours that college students spend preparing for final exams. A national study stated that students at a 4-year college spend an average of 23 hours preparing for 5 final exams each semester with a population standard deviation of 7.3 hours. The researcher sampled 227 students and found a sample mean study time of 19.6 hours. Does this indicate that the average study time for final exams has decreased? Use a 1% level of significance to test this claim.

Step (1) State the null and alternative hypotheses.

$$H_1: \mu = 23 \text{ hours}$$

$$H_2: \mu < 23 \text{ hours}$$

Step (2) State the level of significance and the critical value.

This is a left-sided test so alpha (0.01) is all in the left tail.

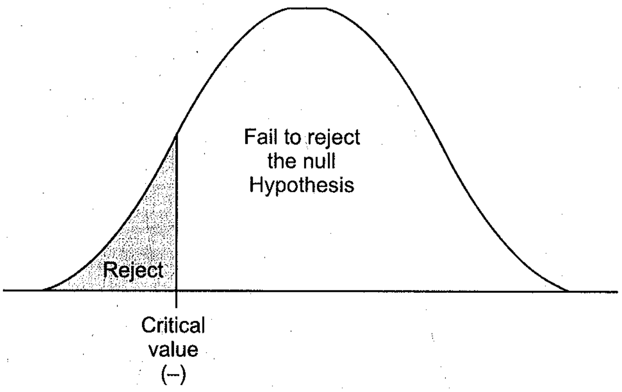


Fig. 5.9 The rejection zone for a left-sided hypothesis test

Go into the body of values in the standard normal table and find the *Z*-score that defines the lower 1% of the area.

The critical value is –2.33. This value sets up the rejection zone.

Step (3) Compute the test statistic.

The test statistic is the number of standard deviations the sample mean is from the known mean. It is also a *Z*-score, just like the critical value.

$$Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

For this problem, the test statistic is

$$Z = \frac{1.96 - 23}{7.3 / \sqrt{227}} = -7.02$$

Step (4) State a conclusion.

Compare the test statistic to the critical value.

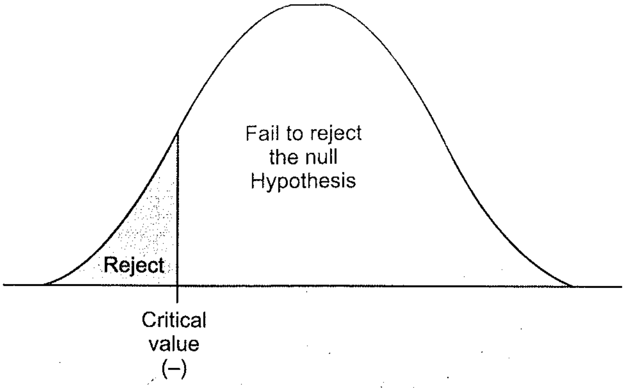


Fig. 5.10 The critical value for a left-sided test when $\alpha = 0.01$

The test statistic falls in the rejection zone. The test statistic of – 7.02 is less than the critical value of – 2.33.

We reject the null hypothesis. We have sufficient evidence to support the claim that the mean final exam study time has decreased below 23 hours.

5.2.2 Testing a Hypothesis using *P*-values

The *p*-value is the probability of observing our sample mean given that the null hypothesis is true. It is the area under the curve to the left or right of the test statistic. If the probability of observing such a sample mean is very small (less than the level of significance), we would reject the null hypothesis. Computations for the *p*-value depends on whether it is a one- or two-sided test.

Steps for a hypothesis test using *p*-values:

- State the null and alternative hypotheses.
- State the level of significance.
- Compute the test statistic and find the area associated with it (this is the *p*-value).
- Compare the *p*-value to alpha (α) and state a conclusion.

Instead of comparing *Z*-score test statistic to *Z*-score critical value, as in the classical method, we compare area of the test statistic to area of the level of significance.

The Decision Rule: If the *p*-value is less than alpha, we reject the null hypothesis

Computing P-values

If it is a two-sided test (the alternative claim is $\neq$), the *p*-value is equal to two times the probability of the absolute value of the test statistic. If the test is a left-sided test (the alternative claim is '<'), then the *p*-value is equal to the area to the left of the test statistic. If the test is a right-sided test (the alternative claim is '>'), then the *p*-value is equal to the area to the right of the test statistic.

Problem 5.4 Problem 5.1 using *p*-value:

A forester studying diameter growth of red pine believes that the mean diameter growth will be different from the known mean growth of 1.35 in./year if a fertilization treatment is applied to the stand. He conducts his experiment, collects data from a sample of 32 plots, and gets a sample mean diameter growth of 1.6 in./year. The population standard deviation for this stand is known to be 0.46 in./year. Does he have enough evidence to support his claim?

Step (1) State the null and alternative hypotheses.

$$H_0: \mu = 1.35 \text{ in./year}$$
$$H_1: \mu \neq 1.35 \text{ in./year}$$

Step (2) State the level of significance.

We will choose a level of significance of 5% ($\alpha = 0.05$).

Step (3) Compute the test statistic.

For this problem, the test statistic is:

$$Z = \frac{1.6 - 1.35}{0.46/\sqrt{32}} = 3.07$$

The *p*-value is two times the area of the absolute value of the test statistic (because the alternative claim is 'not equal').

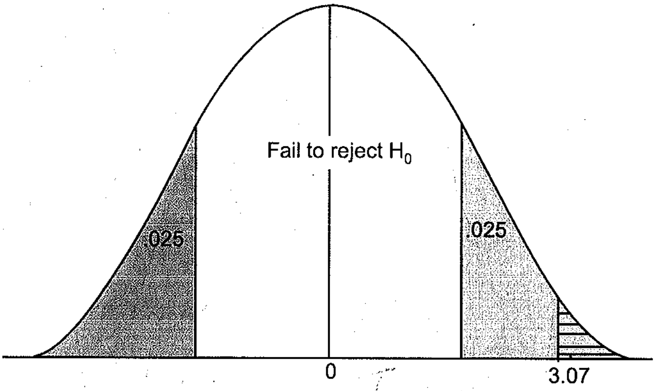


Fig. 5.11 The *p*-value compared to the level of significance

Look up the area for the *Z*-score 3.07 in the standard normal table. The area (probability) is equal to 1 – 0.9989 = 0.0011.

Multiply this by 2 to get the *p*-value = 2 × 0.0011 = 0.0022.

Step (4) Compare the *p*-value to alpha and state a conclusion.

Use the Decision Rule (if the *p*-value is less than α, reject H_0).

In this problem, the *p*-value (0.0022) is less than alpha (0.05).

We reject the H_0. We have enough evidence to support the claim that the mean diameter growth is different from 1.35 inches/year.

Problem 5.5 *Problem 5.2 using p value:*

A researcher believes that there has been an increase in the average farm size in his state since the last study five years ago. The previous study reported a mean size of 450 acres with a population standard deviation (σ) of 167 acres. He samples 45 farms and gets a sample mean of 485.8 acres. Is there enough information to support his claim?

Step (1) State the null and alternative hypotheses.

$$H_2: \mu = 450 \text{ acres}$$
$$H_1: \mu > 450 \text{ acres}$$

Step (2) State the level of significance.

We will choose a level of significance of 5% ($\alpha = 0.05$).

Step (3) Compute the test statistic.

For this problem, the test statistic is

$$Z = \frac{485.8 - 450}{167/\sqrt{45}} = 1.44$$

The p-value is the area to the right of the Z-score 1.44 (the hatched area).

This is equal to $1 - 0.9251 = 0.0749$.

The p-value is 0.0749.

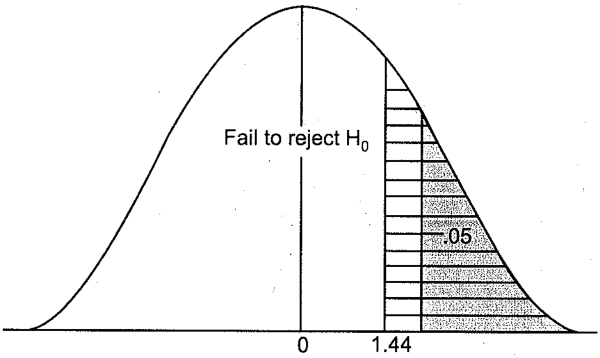


Fig. 5.12 The p-value compared to the level of significance for a right-sided test

Step (4) Compare the p-value to alpha and state a conclusion.

Use the Decision Rule.

In this problem, the p-value (0.0749) is greater than alpha (0.05), so we Fail to Reject the H_0.

The area of the test statistic is greater than the area of alpha (α).

We fail to reject the null hypothesis. We do not have enough evidence to support the claim that the mean farm size has increased.

Problem 5.6 *Problem 5.3 using p value:*

A researcher believes that there has been a reduction in the mean number of hours that college students spend preparing for final exams. A national study stated that students at a 4-year college spend an average of 23 hours preparing for 5 final exams each semester with a population standard deviation of 7.3 hours. The researcher sampled 227 students and found a sample mean study time of 19.6 hours. Does this indicate that the average study time for final exams has decreased? Use a 1% level of significance to test this claim.

Step (1) State the null and alternative hypotheses.

$$H_0: \mu = 23 \text{ hours}$$

$$H_{1:}\ \mu < 23 \text{ hours}$$

Step (2) State the level of significance.

This is a left-sided test so alpha (0.01) is all in the left tail.

Step (3) Compute the test statistic.

For this problem, the test statistic is

$$Z = \frac{1.95 - 23}{7.3/\sqrt{227}} = -7.02$$

The p-value is the area to the left of the test statistic (the little black area to the left of – 7.02). The *Z*-score of –7.02 is not on the standard normal table. The smallest probability on the table is 0.0002. We know that the area for the *Z*-score –7.02 is smaller than this area (probability). Therefore, the *p*-value is < 0.0002.

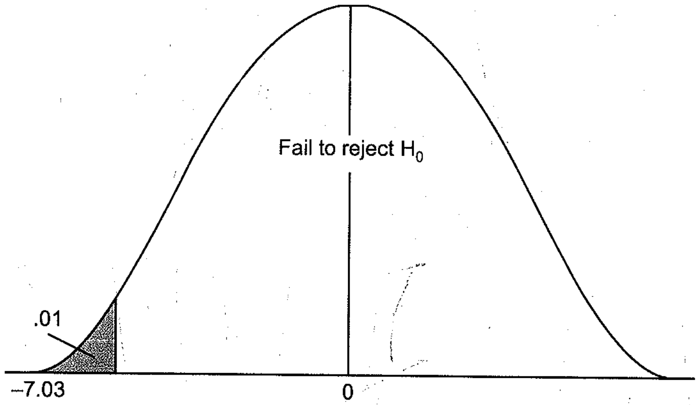


Fig. 5.13 The *p*-value compared to the level of significance for a left-sided test

Step (4) Compare the p-value to alpha and state a conclusion.

Use the Decision Rule.

In this problem, the *p*-value ($p < 0.0002$) is less than alpha (0.01), so we reject the H_0.

The area of the test statistic is much less than the area of alpha (α).

We reject the null hypothesis. We have enough evidence to support the claim that the mean final exam study time has decreased below 23 hours.

Both the classical method and *p*-value method for testing a hypothesis will arrive at the same conclusion. In the classical method, the critical *Z*-score is the number on the z-axis that defines the level of significance (α). The test statistic converts the sample mean to units of standard deviation (a *Z*-score). If the test statistic falls in the rejection zone defined by the critical value, we will reject the null hypothesis. In this approach, two *Z*-scores, which are numbers on the *z*-axis, are compared. In the *p*-value approach, the *p*-value is the area associated with the test statistic. In this method, we compare α (which is also area under the curve) to the *p*-value. If the *p*-value is less than α, we reject the null hypothesis. The *p*-value is the probability of observing such a sample mean when the null hypothesis is true. If the probability is too small (less than the level of significance), then we believe we have enough statistical evidence to reject the null hypothesis and support the alternative claim.

5.3 HYPOTHESIS TEST ABOUT THE POPULATION MEAN (μ) WHEN THE POPULATION STANDARD DEVIATION (σ) IS UNKNOWN

Frequently, the population standard deviation (σ) is not known. We can estimate the population standard deviation (σ) with the sample standard deviation (s). However, the test statistic will no longer follow the standard normal distribution. We must rely on the student's *t*-distribution with $n - 1$ degrees of freedom. Because we use the sample standard deviation (s), the test statistic will change from a *Z*-score to a *t*-score.

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \rightarrow t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

- Steps for a hypothesis test are the same that we covered in Section 5.2.
- State the null and alternative hypotheses.
- State the level of significance and the critical value.
- Compute the test statistic.
- State a conclusion.

Just as with the hypothesis test from the previous section, the data for this test must be from a random sample and requires either that the population from which the sample was drawn be normal or that the sample size is sufficiently large ($n \geq 30$). A *t*-test is robust, so small departures from normality will not adversely affect the results of the test. That being said, if the sample size is smaller than 30, it is always good to verify the assumption of normality through a normal probability plot.

We will still have the same three pairs of null and alternative hypotheses and we can still use either the classical approach or the *p*-value approach.

Selecting the correct critical value from the student's *t*-distribution table depends on three factors: the type of test (one-sided or two-sided alternative hypothesis), the sample size, and the level of significance.

For a two-sided test ('not equal' alternative hypothesis), the critical value ($t_{\alpha/2}$), is determined by alpha (α), the level of significance, divided by two, to deal with the possibility that the result could be less than OR greater than the known value.

If your level of significance was 0.05, you would use the 0.025 column to find the correct critical value ($0.05/2 = 0.025$). If your level of significance was 0.01, you would use the 0.005 column to find the correct critical value ($0.01/2 = 0.005$).

For a one-sided test ('a less than' or 'greater than' alternative hypothesis), the critical value (t_α), is determined by alpha (α), the level of significance, being all in the one side.

If your level of significance was 0.05, you would use the 0.05 column to find the correct critical value for either a left or right-side question. If you are asking a 'less than' (left-sided question, your critical value will be negative. If you are asking a 'greater than' (right-sided question), your critical value will be positive.

5.3.1 Testing a Hypothesis Using Classical Approach

The classical approach is based on standard deviations. This method compares the test statistic (*t*-score) to a critical value (*t*-score) from the student *t*-distribution table. If the test statistic falls in the rejection zone, you reject the null hypothesis.

Problem 5.7:

Find the critical value you would use to test the claim that $\mu \neq 112$ with a sample size of 18 and a 5% level of significance.

In this case, the critical value ($t_{\alpha/2}$) would be 2.110. This is a two-sided question (#) so you would divide alpha by 2 ($0.05/2 = 0.025$) and go down the 0.025 column to 17 degrees of freedom.ample10

What would the critical value be if you wanted to test that $\mu < 112$ for the same data?

In this case, the critical value would be 1.740. This is a one-sided question (<) so alpha would be divided by 1 ($0.05/1 = 0.05$). You would go down the 0.05 column with 17 degrees of freedom to get the correct critical value.

Problem 5.8: *A Two-sided t-Test*

In 2005, the mean pH level of rain in a county in northern New York was 5.41. A biologist believes that the rain acidity has changed. He takes a random sample of 11 rain dates in 2010 and obtains the following data. Use a 1% level of significance to test his claim.

4.70, 5.63, 5.02, 5.78, 4.99, 5.91, 5.76, 5.54, 5.25, 5.18, 5.01

The sample size is small and we don't know anything about the distribution of the population, so we examine a normal probability plot. The distribution looks normal so we will continue with our test.

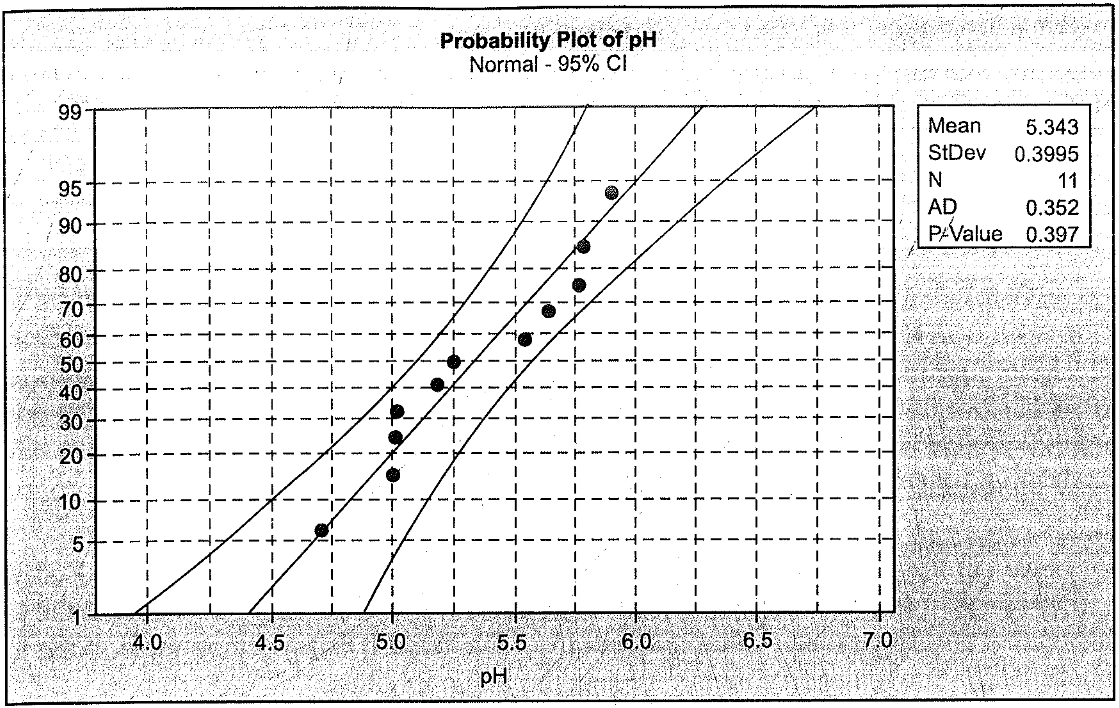


Fig. 5.14 A normal probability plot for Example 9

The sample mean is 5.343 with a sample standard deviation of 0.397.

Step (1) State the null and alternative hypotheses.

$$H_0: \mu = 5.41$$

$$H_1: \mu \neq 5.41$$

Step (2) State the level of significance and the critical value.

This is a two-sided question so alpha is divided by two.

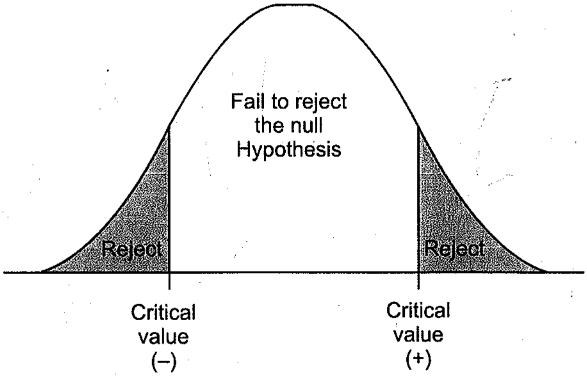

Fail to reject
the null
Hypothesis
Reject
Reject
Critical
value
(–)
Critical
value
(+)

Fig. 5.15 The rejection zones for a two-sided test

$t_{\alpha/2}$ is found by going down the 0.005 column with 14 degrees of freedom.

$$t_{\alpha/2} = \pm 3.269$$

Step (3) Compute the test statistic.

The test statistic is a t-score.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

For this problem, the test statistic is

$$t = \frac{5.343 - 5.41}{0.397/\sqrt{11}} = -0.560$$

Step (4) State a conclusion.

Compare the test statistic to the critical value.

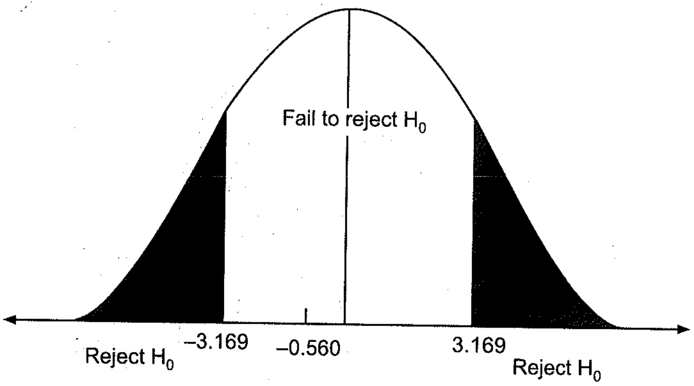


Fig. 5.16 The critical values for a two-sided test when $\alpha = 0.01$

The test statistic does not fall in the rejection zone.

We will fail to reject the null hypothesis. We do not have enough evidence to support the claim that the mean rain pH has changed.

Problem 5.9 *A One-sided t-Test*

Cadmium, a heavy metal, is toxic to animals. Mushrooms, however, are able to absorb and accumulate cadmium at high concentrations. The government has set safety limits for cadmium in dry vegetables at 0.5 ppm. Biologists believe that the mean level of cadmium in mushrooms growing near strip mines is greater than the recommended limit of 0.5 ppm, negatively impacting the animals that live in this ecosystem. A random sample of 51 mushrooms gave a sample mean of 0.59 ppm with a sample standard deviation of 0.29 ppm. Use a 5% level of significance to test the claim that the mean cadmium level is greater than the acceptable limit of 0.5 ppm.

The sample size is greater than 30 so we are assured of a normal distribution of the means.

Step (1) State the null and alternative hypotheses.

$$H_0: \mu = 0.5 \text{ ppm}$$
$$H_1: \mu > 0.5 \text{ ppm}$$

Step (2) State the level of significance and the critical value.

This is a right-sided question so alpha is all in the right tail.

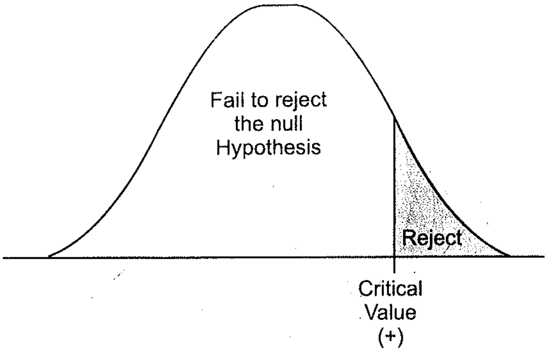

Fail to reject
the null
Hypothesis
Reject
Critical
Value
(+)

Fig. 5.17 Rejection zone for a right-sided test

t_α is found by going down the 0.05 column with 50 degrees of freedom.

$$t_\alpha = 1.676$$

Step (3) Compute the test statistic.

The test statistic is a t-score.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

For this problem, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{0.59 - 0.50}{0.29/\sqrt{51}} = 2.216$$

Step (4) State a Conclusion.

Compare the test statistic to the critical value.

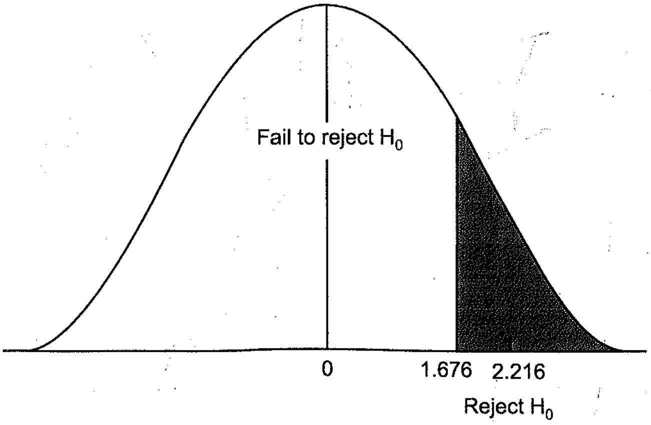

Fail to reject H_0
0
1.676
2.216
Reject H_0

Fig. 5.18 Critical value for a right-sided test when $\alpha = 0.05$

The test statistic falls in the rejection zone. We will reject the null hypothesis. We have enough evidence to support the claim that the mean cadmium level is greater than the acceptable safe limit.

BUT, what happens if the significance level changes to 1%?

The critical value is now found by going down the 0.01 column with 50 degrees of freedom. The critical value is 2.403. The test statistic is now LESS THAN the critical value. The test statistic does not fall in the rejection zone. The conclusion will change. We do NOT have enough evidence to support the claim that the mean cadmium level is greater than the acceptable safe limit of 0.5 ppm.

The level of significance is the probability that you, as the researcher, set to decide if there is enough statistical evidence to support the alternative claim. It should be set before the experiment begins.

5.3.2 *P*-value Approach

We can also use the *p*-value approach for a hypothesis test about the mean when the population standard deviation (σ) is unknown. However, when using a student's *t*-table, we can only estimate the range of the p-value, not a specific value as when using the standard normal table. The student's *t*-table has area (probability) across the top row in the table, with *t*-scores in the body of the table.

To find the *p*-value (the area associated with the test statistic), you would go to the row with the number of degrees of freedom.

Go across that row until you find the two values that your test statistic is between, then go up those columns to find the estimated range for the p-value.

Estimating *P*-value from a Student's *T*-table.

Table 5.3 Portion of the student's *t*-table

t-Distribution

Area in Right Tail

df	.05	.025	.02	.01	.005
1	6.314	12.706	15.894	31.821	63.657
2	2.920	4.303	4.819	6.965	9.925
3	2.353	3.182	3.482	4.541	5.841
4	2.132	2.776	2.999	3.747	4.604
5	2.015	2.571	2.757	3.365	4.032

If your test statistic is 3.789 with 3 degrees of freedom, you would go across the 3 df row. The value 3.789 falls between the values 3.482 and 4.541 in that row. Therefore, the p-value is between 0.02 and 0.01. The p-value will be greater than 0.01 but less than 0.02 ($0.01 < p < 0.02$).

Conclusion

If your level of significance is 5%, you would reject the null hypothesis as the *p*-value (0.01 – 0.02) is less than alpha (α) of 0.05.

If your level of significance is 1%, you would fail to reject the null hypothesis as the p-value (0.01 – 0.02) is greater than alpha (α) of 0.01.

Software packages typically output *p*-values. It is easy to use the Decision Rule to answer your research question by the p-value method.

Excel does not offer 1-sample hypothesis testing.

5.4 HYPOTHESIS TEST FOR A POPULATION PROPORTION (*P*)

Frequently, the parameter we are testing is the population proportion.

We are studying the proportion of trees with cavities for wildlife habitat.

We need to know if the proportion of people who support green building materials has changed.

Has the proportion of wolves that died last year in Yellowstone increased from the year before?

Recall that the best point estimate of *p*, the population proportion, is given by

$$\hat{p} = \frac{x}{n}$$

where x is the number of individuals in the sample with the characteristic studied and n is the sample size. The sampling distribution of $\hat{p}$ is approximately normal with a mean $\mu_\beta = p$ and a standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

when $np(1-p) \geq 10$. We can use both the classical approach and the p-value approach for testing.

The steps for a hypothesis test are the same that we covered in Section 5.2.

- State the null and alternative hypotheses.
- State the level of significance and the critical value.
- Compute the test statistic.
- State a conclusion.

The test statistic follows the standard normal distribution. Notice that the standard error (the denominator) uses p instead of $\hat{p}$, which was used when constructing a confidence interval about the population proportion. In a hypothesis test, the null hypothesis is assumed to be true, so the known proportion is used.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

The critical value comes from the standard normal table, just as in Section 2. We will still use the same three pairs of null and alternative hypotheses as we used in the previous sections, but the parameter is now p instead of μ:

Table 5.2 Different type of Hypotheses

Two sided	Left-sided	Right-sided
$H_0: p = c$ $H_1: p \neq c$	$H_0: p = c$ $H_1: p < c$	$H_0: p = c$ $H_1: p > c$

For a two-sided test, alpha will be divided by 2 giving a $\pm Z\alpha_{/2}$ critical value.

For a left-sided test, alpha will be all in the left tail giving a $-Z_\alpha$ critical value.

For a right-sided test, alpha will be all in the right tail giving a Z_α critical value.xPample 14

Problem 5.10 A botanist has produced a new variety of hybrid soy plant that is better able to withstand drought than other varieties. The botanist knows the seed germination for the parent plants is 75%, but does not know the seed germination for the new hybrid. He tests the claim that it is different from the parent plants. To test this claim, 450 seeds from the hybrid plant are tested and 321 have germinated. Use a 5% level of significance to test this claim that the germination rate is different from 75%.

Step (1) State the null and alternative hypotheses.

$$H_{0:}\ p = 0.75$$
$$H_{1:}\ p \neq 0.75$$

Step (2) State the level of significance and the critical value.

This is a two-sided question so alpha is divided by 2.

Alpha is 0.05 so the critical values are $\pm Z_{\alpha/2} = \pm Z_{\cdot 025}$.

Look on the negative side of the standard normal table, in the body of values for 0.025.

The critical values are ± 1.96.

Step (3) Compute the test statistic.

The test statistic is the number of standard deviations the sample mean is from the known mean. It is also a Z-score, just like the critical value.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

For this problem, the test statistic is

$$z = \frac{0.713 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{450}}} = -1.81$$

Step (4) State a conclusion.

Compare the test statistic to the critical value.

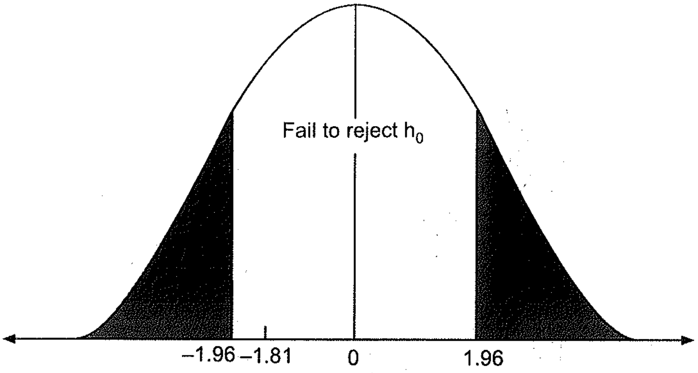


Fig. 5.19. Critical values for a two-sided test when $\alpha = 0.05$.

The test statistic does not fall in the rejection zone. We fail to reject the null hypothesis. We do not have enough evidence to support the claim that the germination rate of the hybrid plant is different from the parent plants.

Let's answer this question using the p-value approach. Remember, for a two-sided alternative hypothesis ("not equal"), the p-value is two times the area of the test statistic. The test statistic is – 1.81 and we want to find the area to the left of –1.81 from the standard normal table.

On the negative page, find the Z-score -1.81. Find the area associated with this Z-score.

The area = 0.0351.

This is a two-sided test so multiply the area times 2 to get the *p*-value = 0.0351 × 2 = 0.0702.

Now compare the p-value to alpha. The Decision Rule states that if the *p*-value is less than alpha, reject the H_0. In this case, the *p*-value (0.0702) is greater than alpha (0.05) so we will fail to reject H_0. We do not have enough evidence to support the claim that the germination rate of the hybrid plant is different from the parent plants.

Problem 5.11 You are a biologist studying the wildlife habitat in the Monongahela National Forest. Cavities in older trees provide excellent habitat for a variety of birds and small mammals. A study five years ago stated that 32% of the trees in this forest had suitable cavities for this type of wildlife. You believe that the proportion of cavity trees has increased. You sample 196 trees and find that 79 trees have cavities. Does this evidence support your claim that there has been an increase in the proportion of cavity trees?

Use a 10% level of significance to test this claim.

Step (1) State the null and alternative hypotheses.

$$H_{0:}\ p = 0.32$$

$$H_{1:}\ p > 0.32$$

Step (2) State the level of significance and the critical value.

This is a one-sided question so alpha is divided by 1.

Alpha is 0.10 so the critical value is $Z_{\alpha} = Z_{.10}$

Look on the positive side of the standard normal table, in the body of values for 0.90.

The critical value is 1.28.

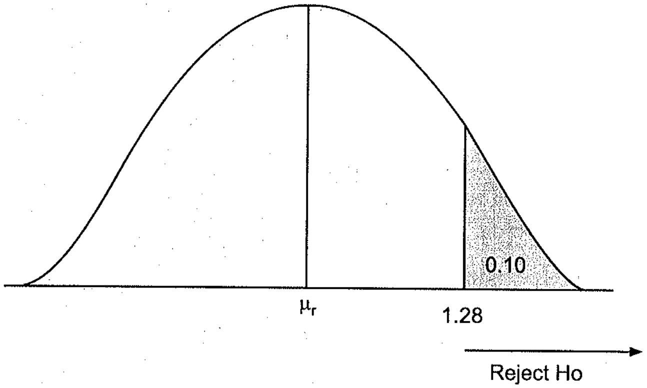


Fig. 5.20 Critical value for a right-sided test where $\alpha = 0.10$

Step (3) Compute the test statistic.

The test statistic is the number of standard deviations the sample proportion is from the known proportion. It is also a Z-score, just like the critical value.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

For this problem, the test statistic is:

$$z = \frac{0.403 - 0.32}{\sqrt{\frac{0.32(1-0.32)}{196}}} = 2.49$$

Step (4) State a conclusion.

Compare the test statistic to the critical value.

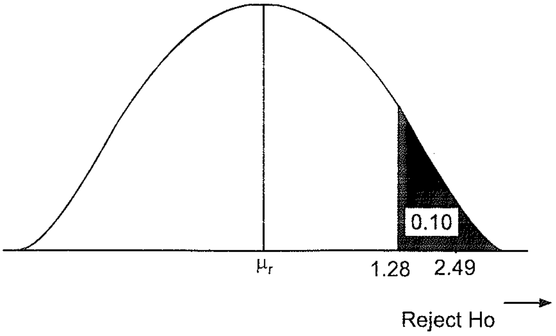


Fig. 5.21 Comparison of the test statistic and the critical value

The test statistic is larger than the critical value (it falls in the rejection zone). We will reject the null hypothesis. We have enough evidence to support the claim that there has been an increase in the proportion of cavity trees.

Now use the *p*-value approach to answer the question. This is a right-sided question ("greater than"), so the *p*-value is equal to the area to the right of the test statistic. Go to the positive side of the standard normal table and find the area associated with the *Z*-score of 2.49. The area is 0.9936. Remember that this table is cumulative from the left. To find the area to the right of 2.49, we subtract from one.

$$p\text{-value} = (1 - 0.9936) = 0.0064$$

The *p*-value is less than the level of significance (0.10), so we reject the null hypothesis. We have enough evidence to support the claim that the proportion of cavity trees has increased.

5.5 HYPOTHESIS TEST ABOUT A VARIANCE

When people think of statistical inference, they usually think of inferences involving population means or proportions. However, the particular population parameter needed to answer an experimenter's practical questions varies from one situation to another, and sometimes a population's variability is more important than its mean. Thus, product quality is often defined in terms of low variability.

Sample variance s^2 can be used for inferences concerning a population variance. For a random sample of *n* measurements drawn from a normal population with mean μ and variance, the value s^2 provides a **point estimate** for. In addition, the quantity follows a **Chi-square (χ^2) distribution**, with $df = n - 1$.

The properties of **Chi-square (χ^2)** distribution are:

Unlike *Z* and *t* distributions, the values in a chi-square distribution are all positive.

The chi-square distribution is asymmetric, unlike the *Z* and *t* distributions.

There are many chi-square distributions. We obtain a particular one by specifying the degrees of freedom ($df = n - 1$) associated with the sample variances s^2.

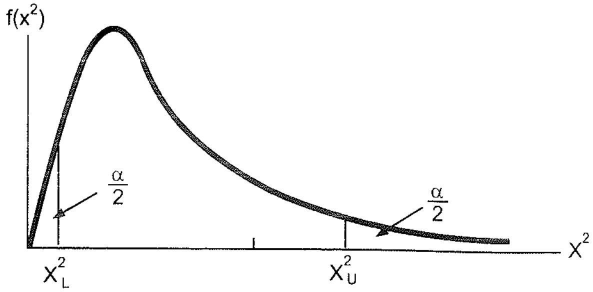


Fig. 5.22. The chi-square distribution.

5.5.1 One-sample Test for Testing the Hypotheses

Null hypothesis: H_0: $\sigma^2 = \sigma_0^2$ (constant)

Alternative hypothesis:

H_a: $\sigma^2 > \sigma_0^2 >$ (one-tailed), reject H_0 if the observed $\chi^2 > \chi_U^2$ (upper-tail value at α).

H_a: $\sigma^2 < \sigma_0^2 <$ (one-tailed), reject H_0 if the observed $\chi^2 < \chi_L^2$ (lower-tail value at α).

H_a: $\neq \sigma_0^2$ (two-tailed), reject H_0 if the observed $\chi^2 > \chi_U^2$ or $\chi^2 < \chi_L^2$ at $\alpha/2$.

where the χ^2 critical value in the rejection region is based on degrees of freedom $df = n - 1$ and a specified significance level of α.

Test statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

As with previous sections, if the test statistic falls in the rejection zone set by the critical value, you will reject the null hypothesis.

Problem 5.11 A forester wants to control a dense understory of striped maple that is interfering with desirable hardwood regeneration using a mist blower to apply an herbicide treatment. She wants to make sure that treatment has a consistent application rate, in other words, low variability not exceeding 0.25 gal./acre (0.06 gal.2). She collects sample data ($n = 11$) on this type of mist blower and gets a sample variance of 0.064 gal.2. Using a 5% level of significance, test the claim that the variance is significantly greater than 0.06 gal.2.

$$H_0: \sigma^2 = 0.06$$
$$H_1: \sigma^2 > 0.06$$

The critical value is 18.307. Any test statistic greater than this value will cause you to reject the null hypothesis.

The test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(11-1)0.064}{0.06} = 10.667$$

We fail to reject the null hypothesis. The forester does NOT have enough evidence to support the claim that the variance is greater than 0.06 gal.2 You can also estimate the *p*-value using the same method as for the student *t*-table. Go across the row for degrees of freedom until you find the two values that your test statistic falls between. In this case going across the row 10, the two table values are 4.865 and 15.987. Now go up those two columns to the top row to estimate the p-value (0.1 – 0.9). The p-value is greater than 0.1 and less than 0.9. Both are greater than the level of significance (0.05) causing us to fail to reject the null hypothesis.

5.6 INFERENCES ABOUT THE DIFFERENCES OF TWO POPULATIONS

Up to this point, we have discussed inferences regarding a single population parameter (e.g., μ, p, σ^2). We have used sample data to construct confidence intervals to estimate the population mean or proportion and to test hypotheses about the population mean and proportion. In both of these chapters, all the examples involved the use of one sample to form an inference about one population. Frequently, we need to compare two sets of data, and make inferences about two populations. This section deals with inferences about two means, proportions, or variances. For example:

- You are studying turkey habitat and want to see if the mean number of brood hens is different in New York compared to Pennsylvania.
- You want to determine if the treatment used in Skaneateles Lake has reduced the number of milfoil plants over the last three years.
- Is the proportion of people who support alternative energy in California greater compared to New York?
- Is the variability in application different between two mist blowers?

These questions can be answered by comparing the differences of:

- Mean number of hens in NY to the mean number of hens in PA.
- Number of plants in 2007 to the number of plants in 2010.
- Proportion of people in CA to the proportion of people in NY.
- Variances between the mist blowers.

5.6.1 Inferences about Two Means with Independent Samples (Assuming Unequal Variances)

Using independent samples means that there is no relationship between the groups. The values in one sample have no association with the values in the other sample. For example, we want to see if the mean life span for hummingbirds in South Carolina is different from the mean life span in North Carolina. These populations are not related, and the samples are independent. We look at the difference of the independent means.

With a two-sample t-test, we compare the population means to each other and again look at the difference. We expect that $\bar{x}_1 - \bar{x}_2$ would be close to $\mu_1 - \mu_2$. The test statistic will use both sample means, sample standard deviations, and sample sizes for the test.

- For a one-sample t-test we used $\frac{s}{\sqrt{n}}$ as a measure of the standard deviation (the standard error).
- We can rewrite $\frac{s}{\sqrt{n}}$ as $\frac{s^2}{n}$.
- The numerator of the test statistic will be $(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)$.
- This has a standard deviation of $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

A two-sample t-test follows the same four steps we saw in one sample t-test.

- Write the null and alternative hypotheses.
- State the level of significance and find the critical value. The critical value, from the student's t-distribution, has the lesser of $n_1 - 1$ and $n_2 - 1$ degrees of freedom.
- Compute the test statistic.
- Compare the test statistic to the critical value and state a conclusion.

The assumptions for two samples test are same as of one sample test. Both samples come from independent random samples. The populations must be normally distributed, or both have large enough sample sizes (n_1 and $n_2 \geq 30$). We will also use the same three pairs of null and alternative hypotheses.

Table 5.3: Null and alternative hypotheses.

Two sided	Left-sided	Right-sided
H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$	H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 < \mu_2$	H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 > \mu_2$

Rewriting the null hypothesis of $\mu_1 = \mu_2$ to $\mu_1 - \mu_1 - \mu_2 = 0$, simplifies the numerator. The test statistic is Welch's approximation (Satterthwaite Adjustment) under the assumption that the independent population variances are not equal.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This test statistic follows the student's t-distribution with the degrees of freedom adjusted by

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)}$$

A simpler alternative to determining degrees of freedom when working a problem longhand is to use the lesser of $n_1 - 1$ or $n_2 - 1$ as the degrees of freedom. This method results in a smaller value for degrees of freedom and therefore a larger critical value. This makes the test more conservative, requiring more evidence to reject the null hypothesis.

Problem 5.12 A forester is studying the number of cavity trees in old growth stands in Adirondack Park in northern New York. He wants to know if there is a significant difference between the mean number of cavity trees in the Adirondack Park and the old growth stands in the Monongahela National Forest. He collects two independent random samples from each forest. Use a 5% level of significance to test this claim.

Adirondack Park	**Monongahela Forest**
n_1 = 51 stands	n_2 = 56 stands
$\bar{x}_1 = 39.6$	$\bar{x}_2 = 43.9$
$s_1 = 9.4$	$s_2 = 10.7$

(1) $\boldsymbol{H_0}$: $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$. There is no difference between the two population means.

H_1: $\mu_1 \neq \mu_2$. There is a difference between the two population means.

(2) The level of significance is 5%. This is a two-sided test so alpha is split into two sides. Computing degrees of freedom using the equation above gives 105 degrees of freedom.

$$df = \frac{\left(\frac{9.4^2}{51} + \frac{10.7}{56}\right)^2}{\frac{1}{51-1}\left(\frac{9.4^2}{51}\right)^2 + \frac{1}{56-1}\left(\frac{10.7^2}{56}\right)^2} = 104.9$$

The critical value ($t_{\alpha/2}$), based on 100 degrees of freedom (closest value in the *t*-table), is ± 1.984. Using 50 degrees of freedom, the critical value is ± 2.009.

(3) The test statistic is

$$t = \frac{(x_1 - x_2) - (m_1 - m_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(39.6 - 43.9) - (0)}{\sqrt{\frac{9.4^2}{51} + \frac{10.7^2}{56}}} = -2.213$$

(4) The test statistic falls in the rejection zone.

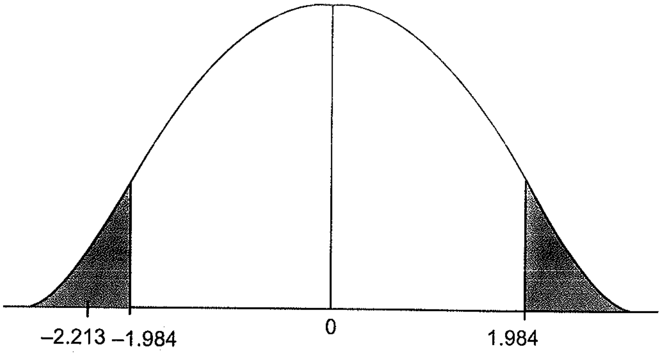


Fig. 5.23 A comparison of the critical values and test statistic

We reject the null hypothesis. We have enough evidence to support the claim that there is a difference in the mean number of cavity trees between the Adirondack Park and the Monongahela National Forest.

5.6.2 Construct and Interpret a Confidence Interval about the Difference of Two Independent Means

A hypothesis test will answer the question about the difference of the means. BUT, we can answer the same question by constructing a confidence interval about the difference of the means.

1. Find the critical value.
2. Compute the margin of error.
3. Point estimate ± margin of error.

Because we are working with two samples, we must modify the components of the confidence interval to incorporate the information from the two populations.

- The point estimate is $\bar{x}_1 - \bar{x}_2$.
- The standard error comes from the test statistic $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- The critical value $t_{\alpha/2}$ comes from the student's *t*-table.

The confidence interval takes the form of the point estimate plus or minus the standard error of the differences.

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We will use the same three steps to construct a confidence interval about the difference of the means.

1. Critical value

2. $E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

3. $\bar{x}_1 - \bar{x}_2 \pm E$

Problem 5.13 Let's look at the mean number of cavity trees in old growth stands again. The forester wants to know if there is a difference between the mean number of cavity trees in old growth stands in the Adirondack forests and in the Monongahela Forest. We can answer this question by constructing a confidence interval about the difference of the means.

1. $t_{\alpha/2} = 2.009$

2. $E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.009\sqrt{\frac{9.4^2}{51} + \frac{10.7^2}{56}} = 3.904$

3. $\bar{x}_1 - \bar{x}_2 \pm 3.904$

The 95% confidence interval for the difference of the means is (– 8.204, – 0.396).

We can be 95% confident that this interval contains the mean difference in number of cavity trees between the two locations. BUT, this doesn't answer the question the forester asked. Is there a difference in the mean number of cavity trees between the Adirondack and Monongahela forests? To answer this, we must look at the confidence interval interpretations.

Confidence Interval Interpretations

- If the confidence interval contains all positive values, we find a significant difference between the groups, AND we can conclude that the mean of the first group is significantly greater than the mean of the second group.
- If the confidence interval contains all negative values, we find a significant difference between the groups, AND we can conclude that the mean of the first group is significantly less than the mean of the second group.
- If the confidence interval contains zero (it goes from negative to positive values), we find NO significant difference between the groups.

In this problem, the confidence interval is (–8.204, –0.396). We have all negative values, so we can conclude that there is a significant difference in the mean number of cavity trees AND that the mean number of cavity trees in the Adirondack forests is significantly less than the mean number of cavity trees in the Monongahela Forest. The confidence interval gives an estimate of the mean difference in number of cavity trees between the two forests. There are, on average, 0.396 to 8.204 fewer cavity trees in the Adirondack Park than the Monongahela Forest.

5.6.3 *P*-value Approach

We can also use the *p*-value approach to answer the question. Remember, the *p*-value is the area under the normal curve associated with the test statistic. This example is a two-sided test (H_1: $\mu_1 \neq \mu_2$) so the *p*-value, when computed by hand, will be multiplied by two.

The test statistic equals –2.213, so the *p*-value is two times the area to the left of –2.213. We can only estimate the *p*-value using the student's *t*-table. Using the lesser of or as the degrees of freedom, we have 50 degrees of freedom. Go across the 50 row in the student's *t*-table until you find the absolute value of the test statistic. In this case, 2.213 falls between 2.109 and 2.403. Going up to the top of each of those columns gives you the estimate of the *p*-value (between 0.02 and 0.01).

Table 5.4 Student *t*-Distribution

Area in Right Tall

df	.05	.025	.02	.01	.005
39	1.686	2.024	2.127	2.429	2.712
40	1.684	2.021	2.123	2.423	2.704
50	1.676	2.009	2.109	2.403	2.678
60	1.671	2.000	2.099	2.390	2.660
70	1.667	1.994	2.093	2.381	2.648

The p-value is 2 × (0.01 – 0.02) = ($0.02 < p < 0.04$). The *p*-value is greater than 0.02 but less than 0.04. This is less than the level of significance (0.05), so we reject the null hypothesis. There is enough evidence to support the claim that there is a significant difference in the mean number of cavity trees between the areas.

PROBLEM 5.14 Researchers are studying the relationship between logging activities in the northern forests and amphibian habitats. They were comparing moisture levels between old-growth and post-harvest habitats. The researchers believe that post-harvest habitat has a lower moisture level. They collected data on moisture levels from two independent random samples. Test their claim using a 5% level of significance.

Old Growth	Post Harvest
$n_1 = 26$	$n_2 = 31$
$\bar{x}_1 = 0.62$ g/cm^3	$\bar{x}_2 = 0.56$ g/cm^3
$s_1 = 0.12$ g/cm^3	$s_2 = 0.17$ g/cm^3

H_0: $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$. There is no difference between the two population means.

H_1: $\mu_1 \neq \mu_2$. Mean moisture level in old growth forests is greater than post-harvest levels.

We will use the critical value based on the lesser of or degrees of freedom. In this problem, there are 25 degrees of freedom and the critical value is 1.708. Now compute the test statistic.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(0.62 - 0.56) - 0}{\sqrt{\frac{0.12^2}{26} + \frac{0.17^2}{31}}} = 1.556$$

The test statistic does not fall in the rejection zone. We fail to reject the null hypothesis. There is not enough evidence to support the claim that the moisture level is significantly lower in the post-harvest habitat.

Now answer this question by constructing a 90% confidence interval about the difference of the means.

1. $t_{\alpha/2} = 1.708$
2. $E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1.708\sqrt{\frac{0.12^2}{26} + \frac{0.17^2}{31}} = 0.0658$
3. $\bar{x}_1 - \bar{x}_2 \pm E\ (0.62 - 0.56) \pm 0.0658$

The 90% confidence interval for the difference of the means is (– 0.0058, 0.1258). The values in the confidence interval run from negative to positive indicating that there is no significant different in the mean moisture levels between old growth and post-harvest stands.

5.6.4 Pooled Two-sampled *t*-test (Assuming Equal Variances)

In the previous section, we made the assumption of unequal variances between our two populations. Welch's *t*-test statistic does not assume that the population variances are equal and can be used whether the population variances are equal or not. The test that assumes equal population variances is referred to as the pooled *t*-test. Pooling refers to finding a weighted average of the two independent sample variances.

The pooled test statistic uses a weighted average of the two sample variances.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \left(\frac{n_1 - 1}{n_1 + n_2 - 2}\right)s_1^2 + \left(\frac{n_2 - 1}{n_1 + n_2 - 2}\right)s_2^2$$

If, $n_1 = n_2$ then $s_p^2 = \frac{1}{2}s_1^2 + \frac{1}{2}s_2^2$, the average of the two sample variances. But whenever $n_1 \neq n_2$, the s^2 based on the larger sample size will receive more weight than the other s^2.

The advantage of this test statistic is that it exactly follows the student's *t*-distribution with $n_1 + n_2 - 2$ degrees of freedom.

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_p^2\left(\frac{1}{n_2} + \frac{1}{n_2}\right)}} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The hypothesis test procedure will follow the same steps as the previous section.

It may be difficult to verify that two population variances might be equal based on sample data. The *F*-test is commonly used to test variances but is not robust. Small departures from normality greatly impact the outcome making the results of the *F*-test unreliable. It can be difficult to decide if a significant outcome from an *F*-test is due to the differences in variances or non-normality. Because of this, many researchers rely on Welch's t when comparing two means.

Problem 5.15 Growth of pine seedlings in two different substrates was measured. We want to know if growth was better in substrate 2. Growth (in cm/yr) was measured and included in the table below. $\alpha = 0.05$

Substrate 1	Substrate 2
3.2	4.5
4.5	6.2
3.8	5.8
4.0	6.0
3.7	7.1
3.2	6.8
4.1	7.2

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 < \mu_2$

$$s_p^2 = \frac{(7-1)\,0.474^2 + (7-1)\,936^2}{7+7-2} = 0.55$$

$$t = \frac{3.79 - 6.23}{\sqrt{0.55\,(1/7 + 1/7)}} = -6.16$$

This is a one-sided test with $n_1 + n_2 - 2 = 12$ degrees of freedom. The critical value is -1.782. The test statistic is less than the critical value so we will reject the null hypothesis.

There is enough evidence to support the claim that the mean growth is less in substrate 1. Growth in substrate 2 is greater.

The confidence interval approach also uses the pooled variance and takes the form:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

using $n_1 + n_2 - 2$ degrees of freedom. So let's answer the same question with a 90% confidence interval.

$$(3.79-6.23) \pm 1.782\sqrt{0.55\,(1/7+1/7)} = (-2.44 \pm 0.7064) = (-3.146, -1.734)$$

All negative values tell you that there is a significant difference between the mean growth for the two substrates and that the growth in substrate 1 is significantly lower than the growth in substrate 2 with reduction in growth ranging from 1.734 to 3.146 cm/yr.

5.6.5 Inferences about Two Means with Dependent Samples–Matched Pairs

Dependent samples occur when there is a relationship between the samples. The data consists of matched pairs from random samples. A sampling method is dependent when the values selected for one sample are used to determine the values in the second sample. Before and after measurements on a population, such as people, lakes, or animals are an example of dependent samples. The objects in your sample are measured twice; measurements are taken at a certain point in time, and then re-taken at a later date. Dependency also occurs when the objects are related, such as eyes or tires on a car. Pairing isn't a problem; it's an opportunity to use the information that occurs with both measurements.

Before you begin your work, you must decide if your samples are dependent. If they are, you can take advantage of this fact. You can use this matching to better answer your research questions. Pairing data reduces measurement variability, which increases the accuracy of our statistical conclusions.

We use the difference (the subtraction) of the pairs of data in our analysis. For each pair, we subtract the values:

- d_1 = before1 – after 1
- d_2 = before 2 – after 2
- d_3 = before 3 – after 3
- ...

We are creating a new random variable d (differences), and it is important to keep the sign, whether positive or negative. We can compute $\bar{d}$, the sample mean of the differences, and s_d, the sample standard deviation of the differences as follows:

$$\bar{d} = \frac{\Sigma d_i}{n}$$

$$s_d = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}}$$

Just as we used the sample mean and the sample standard deviation in a one-sample t-test, we will use the sample mean and sample standard deviation of the differences to test for matched pairs. The assumption of normality must still be verified. The differences must be normally distributed or the sample size must be large enough ($n \geq 30$).

We can do a hypothesis test using matched pairs data following the same methods we used in the previous chapter.

- Write the null and alternative hypotheses.
- State the level of significance and find the critical value.
- Compute a test statistic.
- Compare the test statistic to the critical value and state a conclusion.

Since we are using the differences between the pairs of data, we identify this in our null and alternative hypotheses: H_0: $\mu d = 0$. The mean of the differences is equal to zero; there is no difference in 'before and after' values.

We'll use the same three pairs of null and alternative hypotheses we used in the previous sections.

Table 5.4: Null and alternative hypotheses.

Two sided	Left-sided	Right-sided
H_0: $\mu_d = c$ H_1: $\mu_d \neq c$	H_0: $\mu_d = c$ H_1: $\mu_d < c$	H_0: $\mu_d = c$ H_1: $\mu_d > c$

The critical value comes from the student's t-distribution table with $n - 1$ degrees of freedom, where n = number of matched pairs. The test statistic follows the student's t-distribution.

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

The conclusion must always answer the question you are asking in the alternative hypothesis.

- Reject the H_0. There is enough evidence to support the alternative claim.
- Fail to reject the H_0. There is not enough evidence to support the alternative claim.

Problem 5.16 An environmental biologist wants to know if the water clarity in Owasco Lake is improving. Using a Secchi disk, she takes measurements in specific locations at specific dates during the course of the year. She then repeats the measurements in the same locations and on the same dates five years later. She obtains the following results:

Date	Initial Depth	5-year Depth	Difference
5/11	38	52	–14
6/7	58	60	–2
6/24	65	72	–7

Contd...

7/8	74	72	2
7/27	56	54	2
8/31	36	48	–12
9/30	56	58	–2
10/12	52	60	–8

Using a 5% level of significance, test the biologist's claim that water clarity is improving.

The data are paired by date with two measurements taken at each point five years apart. We will use the differences (right column) to see if there has been a significant improvement in water clarity. Using your calculator, Minitab, or Excel, compute the descriptive statistics on the differences to get the sample mean and the sample standard deviation of the differences.

$$\bar{d} = -5.125$$
$$s_d = 6.081$$

1. The null and alternative hypotheses:

H_0: $\mu_d = 0$ (The mean of the differences is equal to zero- no difference in water clarity over time.)

H_1: $\mu_d < 0$ (The water clarity is improving.)

We test "less than" because of how we computed the differences and the question we are asking.

In this case, we hope to see greater depth (better water clarity) at the five-year measurements. By calculating Initial – 5-year we hope to see negative values, values less than zero, indicating greater depth and clarity at the 5-year mark. Think of it like this:

Initial Depth < 5-year depth

This gives you the direction of the test!

2. The critical value t_α.

The critical value comes from the student's *t*-distribution table with $n - 1$ degrees of freedom. In this problem, we have eight pairs of data ($n = 8$) with 7 degrees of freedom. This is a one-sided test (less than), so alpha is all in the left tail. Go down the 0.05 column with 7 df to find the correct critical value (t_α) of –1.895.

3. The test statistic

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{-5.125 - 0}{6.081/\sqrt{8}} = -2.38$$

We subtract zero from *d*-bar because of our null hypothesis. Our null hypothesis is that the difference of the before and after values are statistically equal to zero. In other words, there has been no change in water clarity.

4. Compare the test statistic to the critical value and state a conclusion.

The test statistic (–2.38) is less than the critical value (–1.895). It falls in the rejection zone.

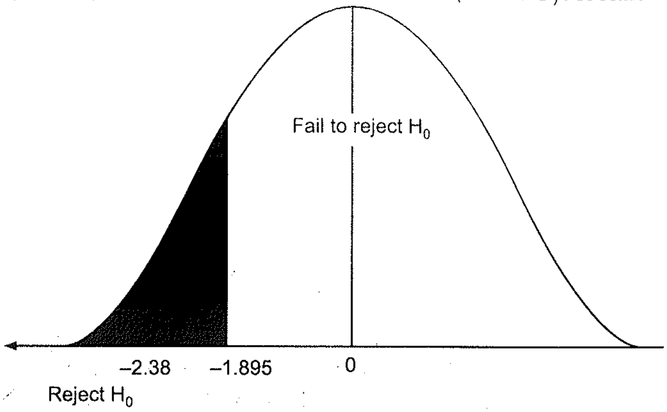


Fig. 5.24 Comparison of the critical value and the test statistic

We reject the null hypothesis. We have enough evidence to support the claim that the mean water clarity has improved.

5.6.6 *P*-value Approach

We can also use the *p*-value approach to answer the question. To estimate *p*-value using the student's *t*-table, go across the row for 7 degrees of freedom until you find the two values that the absolute value of your test statistic falls between.

Table 5.5: Student *t*-Distribution.

df	.05	.025	.02	.01	.005
5	2.015	2.571	2.757	3.365	4.032
6	1.943	2.447	2.612	3.143	3.707
7	1.895	2.365	2.517	2.998	3.499
8	1.860	2.306	2.449	2.896	3.355
9	1.833	2.262	2.398	2.821	3.250

The *p*-value for this test statistic is greater than 0.02 and just less than 0.025. Compare this to the level of significance (alpha). The Decision Rule says that if the *p*-value is less than α, reject the null hypothesis. In this case, the *p*-value estimate (0.02 – 0.025) is less than the level of significance (0.05). Reject the null hypothesis. We have enough evidence to support the claim that the mean water clarity has improved.

BUT, what if you used a 1% level of significance? In this case, the *p*-value is NOT less than the level of significance ((0.02 – 0.025) > 0.01). We would fail to reject the null hypothesis. There is NOT enough evidence to support the claim that the water clarity has improved. It is important to set the level of significance at the start of your research and report the p-value. Another researcher may interpret your findings differently, based on your reported p-value and their own selected level of significance.

5.6.7 Construct and Interpret a Confidence Interval about the Differences of the Data for Matched Pairs

A hypothesis test for matched pairs data is very similar to a one-sample t-test. BUT, we can answer the same question by constructing a confidence interval about the mean of the differences. This process is just like the confidence intervals from Chapter 4.

1. Find the critical value.
2. Compute the margin of error.
3. Point estimate ± margin of error.

For matched pairs data, the critical value comes from the student's t-distribution with $n - 1$ degrees of freedom. The margin of error uses the sample standard deviation of the differences (sd) and the point estimate is $\bar{d}$, the mean of the differences.

For a $(1 - \alpha) \times 100\%$ confidence interval for the mean of the differences

$$\bar{d} \pm t_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right)$$

- Where $t\ \alpha/_2$ is used because confidence intervals are always two-sided.

Problem 5.17 Let's look at the biologist studying water clarity in Owasco Lake again. She wants to test the claim that water clarity has improved. We can answer this question by constructing a confidence interval about the mean of the differences.

$$\bar{d} = -5.125,\ s_d = 6.081,\ \alpha = 0.05,\ n = 8$$

1. $t_{\alpha/2} = 2.365$
2. $E = t_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right) = 2.365\left(\frac{6.081}{\sqrt{8}}\right) = 5.085$
3. $\bar{d} \pm E = -5.125 \pm 5.085$

The 95% confidence interval about the mean of the differences is

$(-10.21, -0.04)$

$(-10.21 \le \mu_d \le -0.04)$

We can be 95% confident that this interval contains the true mean of the differences in water clarity between the two time periods. BUT, this doesn't directly answer the question about improved water clarity. To do this, we use the interpretations given below.

5.6.8 Confidence Interval Interpretations

If the confidence interval contains all positive values, we find a significant difference between the groups, AND we can conclude that the mean of the first group is significantly greater than the mean of the second group.

If the confidence interval contains all negative values, we find a significant difference between the groups, AND we can conclude that the mean of the first group is significantly less than the mean of the second group.

If the confidence interval contains zero (it goes from negative to positive values), we find NO significant difference between the groups.

In this problem, the confidence interval is (–10.21, –0.04). We have all negative values, so we can conclude that there is a significant difference in the mean water clarity between the years AND...

The mean water clarity for the initial time was significantly less than at the five-year re-measurement.

Water clarity has improved during the five-year period. The confidence interval estimates the mean improvement.

Problem 5.18 Biologists are studying elk migration in the western US and want to know if the four-lane interstate that was built ten years ago has disturbed elk migration to the winter feeding area. A random sample was gathered from nine wilderness districts in the winter feeding areas. These data were compared to a random sample collected from the same nine areas before the highway was built. Use a 1% level of significance to test this claim.

District	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Before	11.6	18.7	15.9	20.6	10.1	17.4	7.2	12.2	11.7
After	10.0	21.6	13.9	22.8	11.5	16.2	8.1	10.8	9.6
d	1.6	–2.9	2.0	–2.2	–1.4	1.2	–0.9	1.4	2.1

$\bar{d} = 0.100$ $s_d = 1.946$ $\alpha = 0.01$ $n = 9$

$H_0: \mu_d = 0$

$H_1: \mu_d \neq 0$

Determine the critical values: This is a two-sided question (alternative ≠) so the critical values are ± 3.355.

Compute the test statistic:

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{0.100 - 0}{1.946/\sqrt{9}} = 0.1542$$

Now compare the critical value to the test statistic and state a conclusion. The test statistic is NOT greater than 3.355 or less than –3.355 (it doesn't fall in the rejection zones). We fail to reject the null hypothesis. There is not enough evidence to support the claim that the highway has interfered with the elk migration (no difference before or after the highway).

Now construct a 99% confidence interval and answer the question.

1. $t_{\alpha/2} = 3.355$

2. $E = t_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right) = 3.355\left(\frac{1.946}{\sqrt{9}}\right) = 2.176$

3. $\bar{d} \pm E = 0.100 \pm 2.176$

The 99% confidence interval about the difference of the means is: (–2.076, 2.276)

This confidence interval contains zero. The null hypothesis is that there is zero difference before and after the highway way was created. Therefore, we fail to reject the null hypothesis. There is not enough evidence to support the claim that the highway has interfered with the elk migration (no difference before or after the highway).

5.7 INFERENCES ABOUT TWO POPULATION PROPORTIONS

We can apply the same methods we just learned with means to our two-sample proportion problems. We have two populations with two samples and we want to compare the population proportions.

Is the proportion of lakes in New York with invasive species different from the proportion of lakes in Michigan with invasive species?

Is the proportion of construction companies using certified lumber greater in the northeast than in the southeast?

A test of two population proportions is very similar to a test of two means, except that the parameter of interest is now 'p' instead of 'μ'. With a one-sample proportion test, we used $\hat{p} = \frac{x}{n}$ as the point estimate of p. We expect that $\hat{p}$ would be close to p. With a test of two proportions, we will have two $\tilde{p}$'s, and we expect that $(\tilde{p}_1 - \tilde{p}_2)$ will be close to. The test statistic accounts for both samples.

With a one-sample proportion test, the test statistic is

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

and it has an approximate standard normal distribution.

For a two-sample proportion test, we would expect the test statistic to be

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

HOWEVER, the null hypothesis will be that $p_1 = p_2$. Because the H_0 is assumed to be true, the test assumes that $p_1 = p_2$. We can then assume that $p_1 = p_2$ equals p, a common population proportion. We must compute a pooled estimate of p (its unknown) using our sample data.

$$\tilde{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

The test statistic then takes the form of

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}$$

The hypothesis test follows the same steps that we have seen in previous sections:

- State the null and alternative hypotheses
- State the level of significance and determine the critical value
- Compute the test statistic
- Compare the critical value and the test statistic and state a conclusion

The assumptions that we set for a one-sample proportion test still hold true for both samples. Both must be random samples from normally distributed populations satisfying the following statements:

$$n(p)(1-p) \geq 10$$

Each sample size is no more than 5% of the population size.

We can again use the same three pairs of null and alternative hypotheses. Notice that we are working with population proportions so the parameter is p.

Table 5.5 Null and alternative hypotheses

Two sided	**Left-sided**	**Right-sided**
$H_0: p_1 = p_2$ $H_1: p_1 \neq p_2$	$H_0: p_1 = p_2$ $H_1: p_1 < p_2$	$H_0: p_1 = p_2$ $H_1: p_1 > p_2$

The critical value comes from the standard normal table and depends on the alternative hypothesis (is the question one- or two-sided?). As usual, you must state a conclusion. You must always answer the question that is asked in the alternative hypothesis.

Problem 5.19 A researcher believes that a greater proportion of construction companies in the northeast are using certified lumber in home construction projects compared to companies in the southeast. She collected a random sample of 173 companies in the southeast and found that 86 used at least 30% certified lumber. She collected another random sample of 115 companies from the northeast and found that 68 used at least 30% certified lumber. Test the researcher's claim that a greater proportion of companies in the northeast use at least 30% certified lumber compared to the southeast. $\alpha = 0.05$.

Southeast	**Northeast**
$n_1 = 173$	$n_2 = 115$
$x_1 = 86$	$x_2 = 68$

Write the null and alternative hypotheses:

$$H_0: p_1 = p_2 \text{ or } p_1 - p_2 = 0$$
$$H_1: p_1 < p_2$$

The critical value comes from the standard normal table. It is a one-sided test, so alpha is all in the left tail. The critical value is – 1.645.

Compute the point estimates

$$\hat{p}_1 = \frac{86}{173} = 0.497,\ \hat{p}_2 = \frac{68}{115} = 0.591$$

Now compute $\tilde{p}$

$$\tilde{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{86+68}{173+115} = 0.535$$

The test statistic is

$$z = \frac{(\hat{p}_1 - \hat{p}_2)-(p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{(0.497 - 0.591) - 0}{\sqrt{\frac{0.535(1-0.535)}{173} + \frac{0.535(1-0.535)}{115}}} = -1.57$$

Now compare the critical value to the test statistic and state a conclusion.

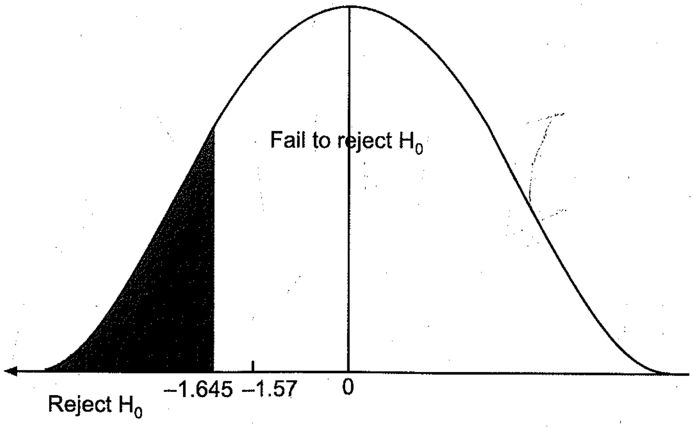


Fig. 5.25 A comparison of the critical value and the test statistic

We fail to reject the null hypothesis. There is not enough evidence to support the claim that a greater proportion of companies in the northeast use at least 30% certified lumber compared to companies in the southeast.

5.7.1 Using the *P*-Value Approach

We can also answer this question using the *p*-value approach. The *p*-value is the area associated with the test statistic. This is a left-tailed problem with a test statistic of –1.57 so the *p*-value is the area to the left of –1.57. Look up the area associated with the *Z*-score –1.57 in the standard normal table.

The *p*-value is 0.0582.

The hatched area (*p*-value) is greater than the 5% level of significance (red area). We fail to reject the null hypothesis. There is not enough statistical evidence to support the claim that a greater proportion of companies in the northeast use at least 30% certified lumber compared to companies in the southeast.

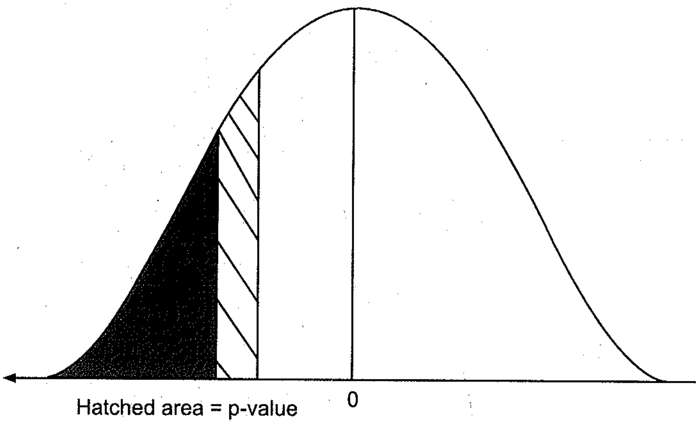


Fig. 5.26 Comparison of p-value and the level of significance

Construct and Interpret a Confidence Interval about the Difference of Two Proportions

Just like a two-sample t-test about the means, we can answer this question by constructing a confidence interval about the difference of the proportions. The point estimate is $\hat{p}_1 - \hat{p}_2$. The standard error is $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ and the critical value comes from the standard normal table.

The confidence interval takes the form of the point estimate ± the margin of error.

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_2)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

We will use the same three steps to construct a confidence interval about the difference of the proportions. Notice the estimate of the standard error of the differences. We do not rely on the pooled estimate of p when constructing confidence intervals to estimate the difference in proportions. This is because we are not making any assumptions regarding the equality of p_1 and p_2, as we did in the hypothesis test.

1. Critical value $z_{\alpha/2}$
2. $E = z_{\alpha/2}\sqrt{\dfrac{p_1(1-p_2)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$
3. $(\hat{p}_1 - \hat{p}_2) \pm E$

Problem 5.20: The researcher claims that a greater proportion of companies in the northeast use at least 30% certified lumber compared to companies in the southeast. We can test this claim by constructing a 90% confidence interval about the difference of the proportions.

1. Critical value $z_{\alpha/2} = 1.645$
2. $E = z_{\alpha/2}\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$

$$= 1.645\sqrt{\frac{0.497(1-0.497}{173} + \frac{0.591(1-0.591)}{115}} = 0.098$$

3. $(\hat{p}_1 - \hat{p}_2) \pm E = (0.497 - 0.591) \pm 0.098$

The 90% confidence interval about the difference of the proportions is (–0.192, 0.004).

BUT, this doesn't answer the question the researcher asked. We must use one of the three interpretations seen in the previous section. In this problem, the confidence interval contains zero. Therefore we can conclude that there is no significant difference between the proportions of companies using certified lumber in the northeast and in the southeast.

Problem 5.21 A hydrologist is studying the use of Best Management Plans (BMP) in managed forest stands to protect riparian zones. He collects information from 62 stands that had a management plan by a forester and finds that 47 stands had correctly implemented BMPs to protect the riparian zones. He collected information from 58 stands that had no management plan and found that 26 of them had correctly implemented BMPs for riparian zones. Do these data suggest that there is a significant difference in the proportion of stands with and without management plans that had correct BMPs for riparian zones? $\alpha = 0.05$.

Plan	**No Plan**
$x_1 = 47$	$x_2 = 26$
$n_1 = 62$	$n_2 = 58$

Let's answer this question both ways by first using a hypothesis test and then by constructing a confidence interval about the difference of the proportions.

$H_0: p_1 = p_2$ or $p_1 - p_2 = 0$

$H_1: p_1 \neq p_2$

Critical value: ± 1.96

Test statistic:

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{(0.758 - 0.448) - 0}{\sqrt{\frac{0.608(1-0.608)}{62} + \frac{0.608(1-0.608)}{58}}} = 3.48$$

The test statistic is greater than 1.96 and falls in the rejection zone. There is enough evidence to support the claim that there is a significant difference in the proportion of correctly implemented BMPs with and without management plans.

Now compute the p-value and compare it to the level of significance. The *p*-value is two times the area under the curve to the right of 3.48. Look for the area (in the standard normal table) associated with a *Z*-score of 3.48. The area to the right of 3.48 is 1 – 0.9997 = 0.0003. The *p*-value is 2 × 0.0003 = 0.0006.

The *p*-value is less than 0.05. We will reject the null hypothesis and support the claim that the proportions are different.

Now, answer this question using a confidence interval.

1. Critical value $z_{\alpha/2} = 1.96$

2. $$E = z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$= 1.96 \times \sqrt{\frac{0.758(1-0.758)}{62} + \frac{0.448(1-0.448)}{58}} = 0.1666$$

3. $(\hat{p}_1 - \hat{p}_2) \pm E = (0.758 - 0.448) \pm 0.1666 = (0.143, 0.477)$

The 95% confidence interval about the difference of the proportions is (0.143, 0.477). The confidence interval contains all positive values, telling you that there is a significant difference between the proportions AND the first group (BMPs used with management plans) is significantly greater than the second group (BMPs with no plans). This confidence interval estimates the difference in proportions. For this problem, we can say that correctly implemented BMPs with a plan occur in a greater proportion (14.3% to 44.7%) compared to those implemented without a management plan.

5.8 F-TEST FOR COMPARING TWO POPULATION VARIANCES

One major application of a test for the equality of two population variances is for checking the validity of the equal variance assumption $(\sigma_1^2 = \sigma_2^2)$ for a two-sample *t*-test. First we hypothesize two populations of measurements that are normally distributed. We label these populations as 1 and 2, respectively. We are interested in comparing the variance of population 1 (σ_1^2) to the variance of population 2 (σ_2^2).

When independent random samples have been drawn from the respective populations, the ratio

$$\frac{s_1^2 / s_2^2}{\sigma_1^2 / \sigma_2^2}$$

possesses a probability distribution in repeated sampling that is referred to as an F distribution and its properties are:

Unlike Z and t, but like χ^2, F can assume only positive values.

The F distribution, unlike the Z and t distributions, but like the χ^2 distribution, is non-symmetrical.

There are many **F** distributions, and each one has a different shape. We specify a particular one by designating the degrees of freedom associated with s_1^2 and s_2^2. We denote these quantities by df_1 and df_2, respectively.

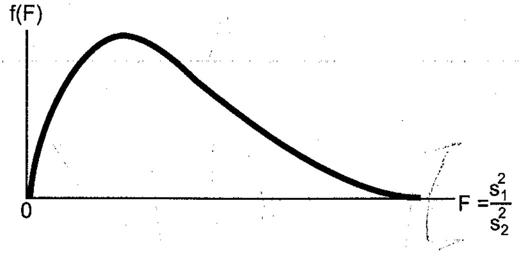


Fig. 5.27 The F-distribution

Note: A statistical test of the null hypothesis $\sigma_1^2 = \sigma_2^2$ utilizes the test statistic s_1^2 / s_2^2. It may require either upper tail or lower tail rejection region, depending on which sample variance is larger. To alleviate this situation, we are at liberty to designate the population with the larger sample variance as **population 1** (i.e., used as the numerator of the ratio). By this convention, the rejection region is only located in the **upper tail of the F distribution.**

Null hypothesis: H_0: $\sigma_1^2 = \sigma_2^2$

Alternative hypothesis:

$H_{a:}\ \sigma_1^2 = \sigma_2^2$ (one-tailed), reject H_0 if the observed $F > F_{\alpha}$.

$H_{a:}\ \sigma_1^2 \quad \sigma_2^2$ two-tailed), reject H_0 if the observed $F > F_{\alpha/2}$.

Test statistic: $F = \dfrac{s_1^2}{s_2^2}$ **assuming** $s_1^2 > s_2^2$,

where the F critical value in the rejection region is based on 2 degrees of freedom $\boldsymbol{df_1 = n_1 - 1}$ (associated with numerator) and $\boldsymbol{df_2 = n_2 - 1}$ (associated with denominator s_2^2).

Problem 5.22 A forester wants to compare two different mist blowers for consistent application. She wants to use the mist blower with the smaller variance, which means more consistent application. She wants to test that the variance of Type A (0.087 gal.2) is significantly greater than the variance of Type B (0.073 gal.2) using $\alpha = 0.05$.

Type A	Type B
$s_1^2 = 0.087$	$s_2^2 = 0.073$
$n_1 = 16$	$n_2 = 21$

$H_{0:}\ \sigma_1^2 = \sigma_2^2$

$H_{1:}\ \sigma_1^2 > \sigma_2^2$

The critical value ($df_1 = 15$ and $df_2 = 20$) is 2.20.

The test statistic is:

$$F = \frac{s_1^2}{s_2^2} = \frac{0.087}{0.073} = 1.192$$

The test statistic is not larger than the critical value (it does not fall in the rejection zone) so we fail to reject the null hypothesis. While the variance of Type B is mathematically smaller than the variance of Type A, it is not statistically smaller. There is not enough statistical evidence to support the claim that the variance of Type A is significantly greater than the variance of Type B. Both mist blowers will deliver the chemical with equal consistency.

Questions about the differences between two samples can be answered in several ways: hypothesis test, *p*-value approach, or confidence interval approach. In all cases, you must clearly state your question, the selected level of significance and the conclusion.

If you choose the hypothesis test approach, you need to compare the critical value to the test statistic. If the test statistic falls in the rejection zone set by the critical value, then you will reject the null hypothesis and support the alternative claim.

If you use the *p*-value approach, you must compute the test statistic and find the area associated with that value. For a two-sided test, the *p*-value is two times the area of the absolute value of the test statistic. For a one-sided test, the *p*-value is the area to the left or right of the test statistic. The decision rule states: If the *p*-value is less than α(level of significance), reject the null hypothesis and support the alternative claim.

The confidence interval approach constructs an interval about the difference of the means or proportions. If the interval contains zero, then you can conclude that there is no difference between the two groups. If the interval contains all positive values, you can conclude that Group 1 is significantly greater than Group 2. If the interval contains all negative numbers, you can conclude that Group 2 is significantly greater than Group 1.

PROBLEMS

1. What is H_0 and H_1 for two-tail test and one-tail test?
2. What is p-value in hypothesis testing?
3. What is the difference between one tail and two tail hypothesis testing?
4. What is the critical value in one tail or two-tail test?
5. Suppose a company has developed a new advertising campaign and claims that it has increased the average sales per customer from \$50 to \$60. To test this claim, a random sample of 40 customers was selected, and their average purchase amount was found to be \$55 with a sample standard deviation of \$8. Assume a significance level of 0.05.
6. Suppose you are investigating whether a new tutoring program has a significant impact on students' exam scores. You collect exam scores from a random sample of 20 students before and after the tutoring program. The data are as follows:

   Before Tutoring: 78, 85, 90, 88, 92, 87, 82, 79, 91, 84, 86, 89, 83, 88, 80, 85, 82, 81, 90, 86

   After Tutoring: 85, 88, 95, 92, 96, 91, 86, 83, 96, 88, 90, 94, 87, 92, 85, 89, 86, 88, 97, 93

   Assuming equal variances, conduct a two-sided t-test at a 5% significance level to determine whether there is a significant difference in exam scores before and after the tutoring program.
7. A weight loss program claims that participants can lose an average of 8 pounds after completing the program. To test this claim, a researcher selects a random sample of 15 participants and records their weight loss in pounds after completing the program. The data are as follows:

   6, 7, 5, 8, 9, 10, 7, 6, 8, 6, 9, 5, 7, 8, 106, 7, 5, 8, 9, 10, 7, 6, 8, 6, 9, 5, 7, 8, 10.

Assuming a normal distribution of weight loss and a significance level of 0.05, conduct a one-sided t-test to determine whether there is evidence that the participants, on average, lose less than 8 pounds after completing the program.

8. A manufacturing company claims that 90% of its products meet a certain quality standard. To test this claim, a random sample of 200 products is selected, and 180 of them meet the quality standard. Conduct a hypothesis test at a 5% significance level to determine whether there is evidence to support the company's claim.

9. Suppose a pharmaceutical company is testing a new drug designed to lower blood pressure. The company claims that their drug is more effective than the currently available drug. To test this claim, a random sample of 30 patients is selected, and their blood pressure reductions after taking each drug are recorded. The data are as follows:

   New Drug: –10, –8, –12, –9, –11, –14, –15, –10, –13, –12, –11, –8, –9, –14, –13, –11, –10, –12, –15, –9, –10, –8, –13, –14, –12, –11, –10, –9, –14, –13.

   Current Drug: –5, –4, –7, –6, –5, –8, –9, –4, –6, –7, –5, –4, -6, –8, –7, –5, –6, –9, –8, –4, –5, –6, –7, –8, –4, –6, –7, –8, –5, –9

   Assume equal variances and a significance level of 0.01. Conduct a hypothesis test to determine whether there is evidence to support the claim that the new drug is more effective in reducing blood pressure than the current drug.

❑❑❑

CHAPTER 6

LINEAR ALGEBRA

Chapter Outline

- Linear Equations • What is a System of Linear Equations? • Solving Linear Systems
- Geometric Interpretation of the Solution Set • Row Reduction and Echelon Forms
- Reduced Row Echelon form (RREF) • Existence and Uniqueness of Solutions

6.1 LINEAR EQUATIONS

Elementary algebra, using the rules of completion and balancing developed by al-Khwarizmi, allows us to determine the value of an unknown variable x that satisfies an equation like the one below:

$$10x - 5 = 15 + 5x$$

An equation like this that only involves an unknown (like x) and not its higher powers (x^2, x^3), along with additions (or subtractions) of the unknown multiplied by numbers (like $10x$ and $5x$) is called a linear equation. We now know, of course, that the equation above can be converted to a special form ('number multiplied by unknown equals number',or a $x = b$, where a and b are numbers):

$$5x = 20$$

Once in this form, it becomes easy to see that $x = b/a = 4$. Linear algebra is, in essence, concerned with the solution of several linear equations in several unknowns.

6.2 WHAT IS A SYSTEM OF LINEAR EQUATIONS?

A system of m linear equations in n unknown variables $x_1, x_2, x_3, \ldots, x_n$ is a collection of m equations of the form

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_3n\, x_n = b_3$$
$$a_{m1}x_1 + a_{m2}x_2 + a_{m3}x_3 + \ldots + a_{mn}x_n = b_m$$

The numbers a_{ij} are called the coefficients of the linear system; because there are m equations and n unknown variables these are therefore $m \times n$ coeffcients. The main problem with a linear system is of course to solve it:

Problem 6.1 Find a list of n numbers $(s_1, s_2, \ldots, s_n)$ that satisfy the above system of linear equations.

In other words, if we substitute the list of numbers $(s_1, s_2, \ldots, s_n)$ for the unknown variables $(x_1, x_2, x_3, \ldots, x_n)$ in the above equation then the left-hand side of the i^{th} equation will be equal to b_i. We call such a list $(s_1, s_2, \ldots, s_n)$ a solution to the system of equations. Note that we say 'a solution' because there may be more than one. The set of all solutions to a linear system is called its solution set. As an example of a linear system, below is a linearsystem consisting of $m = 2$ equations and $n = 3$ unknowns:

$$\begin{aligned} x_1 - 5x_2 - 7x_3 &= 0 \\ 5x_2 + 11x_3 &= 1 \end{aligned}$$

Here is a linear system consisting of $m = 3$ equations and $n = 2$ unknowns:

$$\begin{aligned} -5x_1 + x_2 &= -1 \\ \pi x_1 - \sqrt{5}\, x_2 &= 0 \\ 63x_1 - 2x_2 &= -7 \end{aligned}$$

And finally, below is a linear system consisting of $m = 4$ equations and $n = 6$ unknowns:

$$\begin{aligned} -5x_1 + x_3 - 44x_4 - 55x_6 &= -1 \\ \pi x_1 - 5x_2 - x_3 + 4_{x4} - 5x_5 + \sqrt{5}\, x_6 &= 0 \\ 63x_1 - \sqrt{2}\, x_2 - \frac{1}{5}x_3 + ln(3)x_4 + 4x_5 - \frac{1}{33}x_6 &= 0 \\ 63x_1 - \sqrt{2}x_2 - \frac{1}{5}x_3 - \frac{1}{8}x_4 - 5x_6 &= 5 \end{aligned}$$

Problem 6.2 Verify that $(1, 2, -4)$ is a solution to the system of equations

$$\begin{aligned} 2x_1 + 2x_2 + x_3 &= 2 \\ x_1 + 3x_2 - x_3 &= 11 \end{aligned}$$

Is $(1, -1, 2)$ a solution to the system?

Solution: The number of equations is $m = 2$ and the number of unknowns is $n = 3$. There are $m \times n = 6$ coefficients: $a_{11} = 2$, $a_{12} = 2$, $a_{13} = 1$, $a_{21} = 1$, $a_{22} = 3$, $a_{23} = -1$. And $b_1 = 0$ and $b_2 = 11$. The list of numbers $(1, 2, -4)$ is a solution because

$$\begin{aligned} 2(1) + 2(2) + (-4) &= 2 \\ (1) + 3(2) - (-4) &= 11 \end{aligned}$$

On the other hand, for $(1, -1, 2)$ we have that

$$2(1) + 2(-1) + (2) = 2$$

but $1 + 3(-1) - 2 = -4 = 11$.

Thus, (1, –1, 2) is not a solution to the system.

A linear system may not have a solution at all. If this is the case, we say that the linear system is inconsistent:

INCONSISTENT ⇔ NO SOLUTION

A linear system is called consistent if it has at least one solution:

CONSISTENT ⇔ AT LEAST ONE SOLUTION

We will see shortly that a consistent linear system will have either just one solution or infinitely many solutions. For example, a linear system cannot have just 4 or 5 solutions. If it has multiple solutions, then it will have infinitely many solutions.

Problem 6.3: Show that the linear system does not have a solution.

$$\begin{aligned} -x_1 + x_2 &= 3 \\ x_1 - x_2 &= 1. \end{aligned}$$

Solution. If we add the two equations we get $0 = 4$

which is a contradiction. Therefore, there does not exist a list (s_1, s_2) that satisfies the system because this would lead to the contradiction $0 = 4$.

Problem 6.4: Let t be an arbitrary real number and let

$$s_1 = -\frac{3}{2} - 2t$$

$$s_2 = \frac{3}{2} + t$$

$$s_3 = t$$

Show that for any choice of the parameter t, the list (s_1, s_2, s_3) is a solution to the linear system

$$\begin{aligned} x_1 + x_2 + x_3 &= 0 \\ x_1 + 3x_2 - x_3 &= 3. \end{aligned}$$

Solution. Substitute the list (s_1, s_2, s_3) into the left-hand-side of the first equation

$$\left(-\frac{3}{2} - 2t\right) + \left(\frac{3}{2} + t\right) + (t) = 0$$

And

$$\left(-\frac{3}{2} - 2t\right) + 3\left(\frac{3}{2} + t\right) - (t) = \frac{6}{2} = 3$$

Both equations are satisfied for any value of t. Because we can vary t arbitrarily, we get an in finite number of solutions parameterized by t. For example, compute the list (s_1, s_2, s_3) for $t = 3$ and confirm that the resulting list is a solution to the linear system.

We will use matrices to develop systematic methods to solve linear systems and to study the properties of the solution set of a linear system. Informally speaking, a matrix is an array or table consisting of rows and columns.

For example

$$A = \begin{bmatrix} 1 & -2 & 1 & 0 \\ 0 & 2 & -8 & 8 \\ -4 & 7 & 11 & -5 \end{bmatrix}$$

is a matrix having $m = 3$ rows and $n = 4$ columns. In general, a matrix with m rows and n columns is a $m \times n$ matrix and the set of all such matrices will be denoted by $M_{m \times n}$. Hence, A above is a 3×4 matrix. The entry of A in the i^{th} row and j^{th} column will be denoted by a_{ij}. A matrix containing only one column is called a column vector and a matrix containing only one row is called a row vector. For example, here is a row vector

$$u = [1 \quad -3 \quad 4]$$

and here is a column vector

$$v = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$$

We can associate to a linear system three matrices: (1) the coefficient matrix, (2) the output column vector, and (3) the augmented matrix. For example, for the linear system

$$\begin{aligned} 5x_1 - 3x_2 + 8x_3 &= -1 \\ x_1 + 4x_2 - 6x_3 &= 0 \\ 2x_2 + 4x_3 &= 3 \end{aligned}$$

the coefficient matrix A, the output vector b, and the augmented matrix $[A\ b]$ are:

$$A = \begin{bmatrix} 5 & -3 & 8 \\ 1 & 4 & -6 \\ 0 & 2 & 4 \end{bmatrix}, b = \begin{bmatrix} -1 \\ 0 \\ 3 \end{bmatrix}, [A\ b] = \begin{bmatrix} 5 & -3 & 8 & -1 \\ 1 & 4 & -6 & 0 \\ 0 & 2 & 4 & 3 \end{bmatrix}$$

If a linear system has m equations and n unknowns then the coeffcient matrix A must be a $m \times n$ matrix, that is, A has m rows and n columns. Using our previously defined notation, we can write this as $A \in M_{m \times n}$.

If we are given an augmented matrix, we can write down the associated linear system in an obvious way. For example, the linear system associated to the augmented matrix is

$$\begin{bmatrix} 1 & 4 & -2 & 8 & 12 \\ 0 & 1 & -7 & 2 & -4 \\ 0 & 0 & 5 & -1 & 7 \end{bmatrix}$$

$$\begin{aligned} x_1 + 4x_2 - 2x_3 + 8x_4 &= 12 \\ x_2 - 7x_3 + 2x_4 &= -4 \\ 5x_3 - x_4 &= 7. \end{aligned}$$

We can study matrices without interpreting them as coefficient matrices or augmented matrices associated to a linear system. Matrix algebra is a fascinating subject with numerous applications in every branch of engineering, medicine, statistics, mathematics, finance, biology, chemistry, etc.

6.3 SOLVING LINEAR SYSTEMS

In algebra, you learned to solve equations by first 'simplifying' them using operations that do not alter the solution set. For example, to solve $2x = 8 - 2x$ we can add to both sides $2x$ and obtain $4x = 8$ and then multiply both sides by 1 yielding $x = 2$. We can do similar operations on a linear system. There are three basic operations, called elementary operations, that can be performed:

1. Interchange two equations.
2. Multiply an equation by a nonzero constant.
3. Add a multiple of one equation to another.

These operations do not alter the solution set. The idea is to apply these operations iteratively to simplify the linear system to a point where one can easily write down the solution set. It is convenient to apply elementary operations on the augmented matrix [A b] representing the linear system. In this case, we call the operations elementary row operations, and the process of simplifying the linear system using these operations is called row reduction. The goal with row reducing is to transform the original linear system into one having a triangular structure and then perform back substitution to solve the system. This is best explained via an example.

Problem 6.5 Use back substitution on the augmented matrix

$$\begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

to solve the associated linear system.

Solution. Notice that the augmented matrix has a triangular structure. The third row corresponds to the equation $x_3 = 1$. The second row corresponds to the equation

$x_2 - x_3 = 0$ and therefore $x_2 = x_3 = 1$. The first row corresponds to the equation $x_1 - 2x_3 = -4$ and therefore

$$x_1 = -4 + 2x_3 = -4 + 2 = -2.$$

Therefore, the solution is $(-2, 1, 1)$.

Problem 6.6 Solve the linear system using elementary row operations.

$$\begin{aligned} -3x_1 + 2x_2 + 4x_3 &= 12 \\ x_1 - 2x_3 &= -4 \\ 2x_1 - 3x_2 + 4x_3 &= -3 \end{aligned}$$

Solution. Our goal is to perform elementary row operations to obtain a triangular structure and then use back substitution to solve. The augmented matrix is

$$\begin{bmatrix} -3 & 2 & 4 & 12 \\ 1 & 0 & -2 & -4 \\ 2 & -3 & 4 & -3 \end{bmatrix}$$

Interchange Row 1 (R_1) and Row 2 (R_2):

$$\begin{bmatrix} -3 & 2 & 4 & 12 \\ 1 & 0 & -2 & -4 \\ 2 & -3 & 4 & -3 \end{bmatrix} \xrightarrow[R_1 \leftrightarrow R_2]{} \begin{bmatrix} 1 & 0 & -2 & -4 \\ -3 & 2 & 4 & 12 \\ 2 & -3 & 4 & -3 \end{bmatrix}$$

As you will see, this first operation will simplify the next step. Add $3R_1$ to R_2:

$$\begin{bmatrix} 1 & 0 & -2 & -4 \\ -3 & 2 & 4 & 12 \\ 2 & -3 & 4 & -3 \end{bmatrix} \xrightarrow[3R_1 + R_2]{} \begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 2 & -2 & 0 \\ 2 & -3 & 4 & -3 \end{bmatrix}$$

Add $-2R_1$ to R_3:

$$\begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 2 & -2 & 0 \\ 2 & -3 & 4 & -3 \end{bmatrix} \xrightarrow[-2R_1 + R_3]{} \begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 2 & -2 & 0 \\ 0 & -3 & 8 & 5 \end{bmatrix}$$

Multiply R_2 by $\frac{1}{2}$:

$$\begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 2 & -2 & 0 \\ 0 & -3 & 8 & 5 \end{bmatrix} \xrightarrow[\frac{1}{2}R_2]{} \begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 1 & -1 & 0 \\ 0 & -3 & 8 & 5 \end{bmatrix}$$

Add $3R_2$ to R_3:

$$\begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 1 & -1 & 0 \\ 0 & -3 & 8 & 5 \end{bmatrix} \xrightarrow[3R_2+R_3]{} \begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 5 & 5 \end{bmatrix}$$

Multiply R_3 by $\frac{1}{5}$:

$$\begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 5 & 5 \end{bmatrix} \xrightarrow[\frac{1}{5}R_3]{} \begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

We can continue row reducing but the row reduced augmented matrix is in triangular form. So now use back substitution to solve. The linear system associated to the row reduced augmented matrix is

$$\begin{aligned} x_1 - 2x_3 &= -4 \\ x_2 - x_3 &= 0 \\ x_3 &= 1 \end{aligned}$$

The last equation gives that $x_3 = 1$. From the second equation, we obtain that $x_2 - x_3 = 0$, and thus $x_2 = 1$. The first equation then gives that $x_1 = -4 + 2(1) = -2$. Thus, the solution to the original system is $(-2, 1, 1)$. You should verify that $(-2, 1, 1)$ is a solution to the original system

The original augmented matrix of the previous example is:

$$M = \begin{bmatrix} -3 & 2 & 4 & 12 \\ 1 & 0 & -2 & -4 \\ 2 & -3 & 4 & -3 \end{bmatrix} \rightarrow \begin{aligned} -3x_1 + 2x_2 + 4x_3 &= 12 \\ x_2 - 2x_3 &= -4 \\ 2x_1 - 3x_2 + 4x_3 &= -3 \end{aligned}$$

After row reducing, we obtained the row reduced matrix

$$N = \begin{bmatrix} 1 & 0 & -2 & -4 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \rightarrow \begin{aligned} x_1 + 2x_3 &= -4 \\ x_2 - x_3 &= -0 \\ x_3 &= 1 \end{aligned}$$

Although the two augmented matrices M and N are clearly distinct, it is a fact that they have the same solution set.

Problem 6.7 Using elementary row operations, show that the linear system is inconsistent.

$$\begin{aligned} x_1 + 2x_3 &= 1 \\ x_2 + x_3 &= 0 \\ 2x_1 + 4x_3 &= 1 \end{aligned}$$

Solution. The augmented matrix is

$$\begin{bmatrix} 1 & 0 & 21 \\ 0 & 1 & 10 \\ 2 & 0 & 41 \end{bmatrix}$$

Perform the operation $-2R_1 + R_3$:

$$\begin{bmatrix} 1 & 0 & 21 \\ 0 & 1 & 10 \\ 2 & 0 & 41 \end{bmatrix} \xrightarrow[-2R_1 + R_3]{} \begin{bmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

The last row of the simplified augmented matrix

$$\begin{bmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

corresponds to the equation

$$0\,x_1 + 0\,x_2 + 0\,x_3 = -1$$

Obviously, there are no numbers x_1, x_2, x_3 that satisfy this equation, and therefore, the linear system is inconsistent, *i.e.*, it has no solution. In general, if we obtain a row in an augmented matrix of the form

$$[0\ 0\ 0\ \ldots\ 0\ \ c]$$

where c is a nonzero number, then the linear system is inconsistent. We will call this type of row an inconsistent row. However, a row of the form

$$[0\ \ 1\ \ 0\ \ 0\ \ 0]$$

corresponds to the equation $x_2 = 0$ which is perfectly valid.

6.4 GEOMETRIC INTERPRETATION OF THE SOLUTION SET

The set of points (x_1, x_2) that satisfy the linear system

$$\begin{aligned} x_1 - 2x_2 &= -1 \\ -x_1 + 3x_2 &= 3 \end{aligned}$$

is the intersection of the two lines determined by the equations of the system. The solution for this system is (3, 2). The two lines intersect at the point $(x_1, x_2) = (3, 2)$, see Figure 6.1.

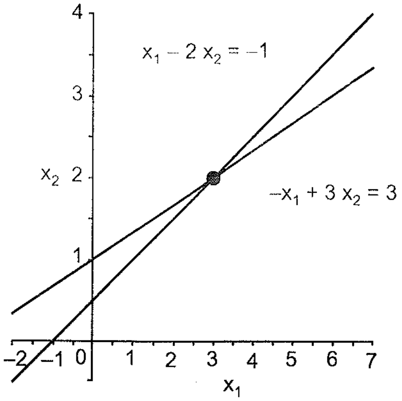


Fig. 6.1

Fig. 6.1 The intersection point of the two lines is the solution of the linear system

Similarly, the solution of the linear system

$$\begin{aligned} x_1 - 2x_2 + x_3 &= 0 \\ 2x_2 - 8x_3 &= 8 \\ -4x_1 + 5x_2 + 9x_3 &= -9 \end{aligned}$$

is the intersection of the three planes determined by the equations of the system. In this case, there is only one solution: (29, 16, 3). In the case of a consistent system of two equations, the solution set is the line of intersection of the two planes determined by the equations of the system, see Fig. 6.2.

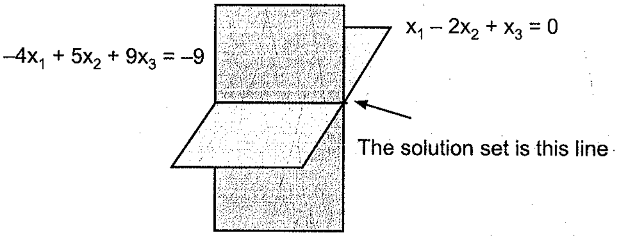


Fig. 6.2

Fig. 6.2 The intersection of the two planes is the solution set of the linear system

6.5 ROW REDUCTION AND ECHELON FORMS

Consider the linear system

$$\begin{aligned} x_1 + 5x_2 - 2x_4 - x_5 + 7x_6 &= -4 \\ 2x_2 - 2x_3 + 3x_6 &= 0 \\ -9x_4 - x_5 + x_6 &= -1 \\ 5x_5 + x_6 &= 5 \\ 0 &= 0 \end{aligned}$$

having augmented matrix

$$\begin{bmatrix} 1 & 5 & 0 & -2 & -1 & 7 & -4 \\ 0 & 2 & -2 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & -9 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 5 & 1 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The above augmented matrix has the following properties:

P1. All nonzero rows are above any rows of all zeros.

P2. The leftmost nonzero entry of a row is to the right of the leftmost nonzero entry of the row above it.

Any matrix satisfying properties P_1 and P_2 is said to be in row echelon form (REF). In REF, the leftmost nonzero entry in a row is called a leading entry:

$$\begin{bmatrix} 1 & 5 & 0 & -2 & -1 & 7 & -4 \\ 0 & 2 & -2 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & -9 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 5 & 1 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

A consequence of property P_2 is that every entry below a leading entry is zero

$$\begin{bmatrix} 1 & 5 & 0 & -2 & -1 & 7 & -4 \\ 0 & 2 & -2 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & -9 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 5 & 1 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We can perform elementary row operations, or row reduction, to transform a matrix into REF.

Problem 6.8 Explain why the following matrices are not in REF. Use elementary row operations to put them in REF.

$$M = \begin{bmatrix} 3 & -1 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 3 & 0 \end{bmatrix} \quad N = \begin{bmatrix} 7 & 5 & 0 & 3 \\ 0 & 3 & -1 & 1 \\ 0 & 6 & -5 & 2 \end{bmatrix}$$

Solution. Matrix M fails property $P1$. To put M in REF we interchange R_2 with R_3:

$$\begin{bmatrix} 3 & -1 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 3 & 0 \end{bmatrix} \xrightarrow[R_2 \leftrightarrow R_3]{} \begin{bmatrix} 3 & -1 & 0 & 3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The above augmented matrix has the following properties:

P1. All nonzero rows are above any rows of all zeros.

P2. The leftmost nonzero entry of a row is to the right of the leftmost nonzero entry of the row above it.

Any matrix satisfying properties P1 and P2 is said to be in row echelon form (REF). In REF, the leftmost nonzero entry in a row is called a leading entry:

$$\begin{bmatrix} 1 & 5 & 0 & -2 & -1 & 7 & -4 \\ 0 & 2 & -2 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & -9 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 5 & 1 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

A consequence of property P2 is that every entry below a leading entry is zero

$$\begin{bmatrix} 1 & 5 & 0 & -2 & -1 & 7 & -4 \\ 0 & 2 & -2 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & -9 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 5 & 1 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We can perform elementary row operations, or row reduction, to transform a matrix into REF.

Problem 6.9 Explain why the following matrices are not in REF. Use elementary row operations to put them in REF.

$$M = \begin{bmatrix} 3 & -1 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 3 & 0 \end{bmatrix} \quad N = \begin{bmatrix} 7 & 5 & 0 & -3 \\ 0 & 3 & -1 & 1 \\ 0 & 6 & -5 & 2 \end{bmatrix}$$

Solution. Matrix M fails property P1. To put M in REF we interchange R_2 with R_3:

$$\begin{bmatrix} 3 & -1 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 3 & 0 \end{bmatrix} \xrightarrow[R_2 \leftrightarrow R_3]{} \begin{bmatrix} 3 & -1 & 0 & 3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The matrix N fails property P2. To put N in REF we perform the operation $-2R_2 + R_3 \rightarrow R_3$:

$$\begin{bmatrix} 7 & 5 & 0 & -3 \\ 0 & 3 & -1 & 1 \\ 0 & 6 & -5 & 2 \end{bmatrix} \xrightarrow[-2R_2+R_3\to R_3]{} \begin{bmatrix} 7 & 5 & 0 & -3 \\ 0 & 3 & -1 & 1 \\ 0 & 0 & -3 & 0 \end{bmatrix}$$

Why is REF useful? Certain properties of a matrix can be easily deduced if it is in REF. For now, REF is useful to us for solving a linear system of equations. If an augmented matrix is in REF, we can use back substitution to solve the system, just as we did in the previous problems.

For example, consider the system

$$\begin{aligned} 8x_1 - 2x_2 + x_3 &= 4 \\ 3x_2 - x_3 &= 7 \\ 2x_3 &= 4 \end{aligned}$$

whose augmented matrix is already in REF.

$$\begin{bmatrix} 8 & -1 & 1 & 4 \\ 0 & 3 & -1 & 7 \\ 0 & 0 & 2 & 4 \end{bmatrix}$$

From the last equation we obtain that $2x_3 = 4$, and thus $x_3 = 2$. Substituting $x_3 = 2$ into the second equation we obtain that $x_2 = 3$. Substituting $x_3 = 2$ and $x_2 = 3$ into the first equation we obtain that $x_1 = 1$.

6.6 REDUCED ROW ECHELON FORM (RREF)

Although REF simplifies the problem of solving a linear system, later in the course we will need to completely row reduce matrices into what is called reduced row echelon form (RREF). A matrix is in RREF if it is in REF (so it satisfies properties P1 and P2) and in addition satisfies the following properties:

P3. The leading entry in each nonzero row is a 1.

P4. All the entries above (and below) a leading 1 are all zero.

A leading 1 in the RREF of a matrix is called a pivot. For example, the following matrix in RREF:

$$\begin{bmatrix} 1 & 6 & 2 & 3 & 0 & 0 \\ 0 & 0 & 1 & -4 & 0 & 5 \\ 0 & 0 & 0 & 0 & 1 & 7 \end{bmatrix}$$

has three pivots:

$$\begin{bmatrix} 0 & 3 & -6 & 6 & 4 & -5 \\ 3 & -7 & 8 & -5 & 8 & 9 \\ 3 & -9 & 12 & -9 & 6 & 15 \end{bmatrix}$$

Problem 6.10 Use row reduction to transform the matrix into RREF.

$$\begin{bmatrix} 0 & 3 & -6 & 6 & 4 & -5 \\ 3 & -7 & 8 & -5 & 8 & 9 \\ 3 & -9 & 12 & -9 & 6 & 15 \end{bmatrix}$$

Solution. The first step is to make the top leftmost entry nonzero:

$$\begin{bmatrix} 0 & 3 & -6 & 6 & 4 & -5 \\ 3 & -7 & 8 & -5 & 8 & 9 \\ 3 & -9 & 12 & -9 & 6 & 15 \end{bmatrix} \xrightarrow[R_3 \leftrightarrow R_1]{} \begin{bmatrix} 3 & -9 & 12 & -9 & 6 & 15 \\ 3 & -7 & 8 & -5 & 8 & 9 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix}$$

Now create a leading 1 in the first row:

$$\begin{bmatrix} 3 & -9 & 12 & -9 & 6 & 15 \\ 0 & -7 & 8 & -5 & 8 & 9 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix} \xrightarrow[R_1 \to \frac{1}{3}R_1]{} \begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 3 & -7 & 8 & -5 & 8 & 9 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix}$$

Create zeros under the newly created leading 1:

$$\begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 3 & -7 & 8 & -5 & 8 & 9 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix} \xrightarrow[3R_i+R_2]{} \begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 0 & 2 & -4 & 4 & 2 & -6 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix}$$

Create a leading 1 in the second row:

$$\begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 0 & 2 & -4 & 4 & 2 & -6 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix} \xrightarrow[\frac{1}{2}R_2]{} \begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 0 & 1 & -2 & 2 & 1 & -3 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix}$$

Create zeros under the newly created leading 1:

$$\begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 0 & 1 & -2 & 2 & 1 & -3 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix} \xrightarrow[-3R_2+R_3]{} \begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 0 & 1 & -4 & 4 & 2 & -6 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix}$$

We have now completed the top-to-bottom phase of the row reduction algorithm. In the next phase, we work bottom-to-top and create zeros above the leading 1's. Create zeros above the leading 1 in the third row:

$$\begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 0 & 1 & -2 & 2 & 1 & -3 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix} \xrightarrow[-R_3+R_2]{} \begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 0 & 1 & -2 & 2 & 1 & -3 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix}$$

$$\begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 0 & 1 & -2 & 2 & 0 & -7 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix} \xrightarrow[-2R_3+R_1]{} \begin{bmatrix} 1 & -3 & 4 & -3 & 0 & -3 \\ 0 & 1 & -2 & 2 & 0 & -7 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix}$$

Create zeros above the leading 1 in the second row:

$$\begin{bmatrix} 1 & -3 & 4 & -3 & 0 & -5 \\ 0 & 1 & -2 & 2 & 0 & -7 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix} \xrightarrow[3R_2+R_1]{} \begin{bmatrix} 1 & 0 & -2 & 3 & 0 & 24 \\ 0 & 1 & -2 & 2 & 20 & -7 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix}$$

This completes the row reduction algorithm, and the matrix is in RREF.

Problem 6.11 Use row reduction to solve the linear system

$$\begin{aligned} 2x_1 + 4x_2 + 6x_3 &= 8 \\ x_1 + 2x_2 + 4x_3 &= 8 \\ 3x_1 + 6x_2 + 9x_3 &= 12 \end{aligned}$$

Solution. The augmented matrix is

$$\begin{bmatrix} 2 & 4 & 6 & 8 \\ 1 & 2 & 4 & 8 \\ 3 & 6 & 9 & 12 \end{bmatrix}$$

Create a leading 1 in the first row:

$$\begin{bmatrix} 2 & 4 & 6 & 8 \\ 1 & 2 & 4 & 8 \\ 3 & 6 & 9 & 12 \end{bmatrix} \xrightarrow[\frac{1}{2}R_1]{} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 8 \\ 3 & 6 & 9 & 12 \end{bmatrix}$$

Create zeros under the first leading 1:

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 8 \\ 3 & 6 & 9 & 12 \end{bmatrix} \xrightarrow[-R_1+R_2]{} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 4 \\ 3 & 6 & 9 & 12 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 4 \\ 3 & 6 & 9 & 12 \end{bmatrix} \xrightarrow[-R_1+R_3]{} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The system is consistent, however, there are only 2 nonzero rows but 3 unknown variables. This means that the solution set will contain $3 - 2 = 1$ free parameter. The second row in the augmented matrix is equivalent to the equation: $x_3 = 4$.

The first row is equivalent to the equation:

$$x_1 + 2x_2 + 3x_3 = 4$$

and after substituting $x_3 = 4$ we obtain

$$x_1 + 2x_2 = -8.$$

We now must choose one of the variables x_1 or x_2 to be a parameter, say t, and solve for the remaining variable. If we set $x_2 = t$ then from $x_1 + 2x_2 = -8$ we obtain that

$$x_1 = -8 - 2t.$$

We can therefore write the solution set for the linear system as

$$\begin{aligned} x_1 &= -8-2t \\ x_2 &= t \\ x_3 &= 4 \end{aligned}$$

where t can be any real number. If we had chosen x_1 to be the parameter, say $x_1 = t$, then the solution set can be written as

$$\begin{aligned} x_1 &= t \\ x_2 &= -4-\frac{1}{2}t \\ x_3 &= 4 \end{aligned}$$

Both are two different parameterizations, they both give the same solution set.

In general, if a linear system has n unknown variables and the row reduced augmented matrix has r leading entries, then the number of free parameters d in the solution set is $d = n - r$. Thus, when performing back substitution, we will have to set d of the unknown variables to arbitrary parameters. In the previous example, there are $n = 3$ unknown variables and the row reduced augmented matrix contained $r = 2$ leading entries. The number of free parameters was therefore $d = n - r = 3 - 2 = 1$. Because the number of leading entries r in the row reduced coefficient matrix determine the number of free parameters, we will refer to r as the rank of the coefficient matrix: $r = rank\ (A)$.

Problem 6.12 Solve the linear system represented by the augmented matrix

$$\begin{bmatrix} 1 & -7 & 2 & -5 & 8 & 10 \\ 0 & 1 & -3 & 3 & 1 & -5 \\ 0 & 0 & 0 & 1 & -1 & 4 \end{bmatrix}$$

Solution. The number of unknowns is $n = 5$ and the augmented matrix has rank $r = 3$ (leading entries). Thus, the solution set is parameterized by $d = 5 - 3 = 2$ free variables, call them t and s. The last equation of the augmented matrix is $x_4 - x_5 = 4$. We choose x_5 to be the first parameter so we set $x_5 = t$. Therefore, $x_4 = 4 + t$. The second equation of the augmented matrix is

$$x_2 - 3x_3 + 3x_4 + x_5 = -5$$

and the unassigned variables are x_2 and x_3. We choose x_3 to be the second parameter, say $x_3 = s$. Then

$$\begin{aligned} x_2 &= -5 + 3x_3 - 3x_4 - x_5 \\ &= -5 + 3s - 3(4 + t) - t \\ &= -17 - 4t + 3s. \end{aligned}$$

We now use the first equation of the augmented matrix to write x_1 in terms of the other variables:

$$\begin{aligned} x_1 &= 10 + 7x_2 - 2x_3 + 5x_4 - 8x_5 \\ &= 10 + 7(-17 - 4t + 3s) - 2s + 5(4 + t) - 8t \\ &= -89 - 31t + 19s \end{aligned}$$

Thus, the solution set is

$$\begin{aligned} x_1 &= -89 - 31t + 19s \\ x_2 &= -17 - 4t + 3s \\ x_3 &= s \\ x_4 &= 4 + t \\ x_5 &= t \end{aligned}$$

where t and s are arbitrary real numbers. Choose arbitrary numbers for t and s and substitute the corresponding list $(x_1, x_2, ..., x_5)$ into the system of equations to verify that it is a solution.

6.7 EXISTENCE AND UNIQUENESS OF SOLUTIONS

The REF or RREF of an augmented matrix leads to three distinct possibilities for the solution set of a linear system.

Let $[A\ b]$ be the augmented matrix of a linear system. One of the following distinct possibilities will occur:

1. The augmented matrix will contain an inconsistent row.
2. All the rows of the augmented matrix are consistent and there are no free parameters.
3. All the rows of the augmented matrix are consistent and there are $d \geq 1$ variables that must be set to arbitrary parameters

In Case 1., the linear system is inconsistent and thus has no solution. In Case 2., the linear system is consistent and has only one (and thus unique) solution. This case occurs when $r = rank(A) = n$ since then the number of free parameters is $d = n - r = 0$. In Case 3., the linear system is consistent and has infinitely many solutions. This case occurs when $r < n$ and thus $d = n - r > 0$ is the number of free parameters.

PROBLEMS

1. Consider the following system of linear equations:

$$\begin{aligned} 2x + y - z &= 5 \\ 4x - 2y + 3z &= 3 \\ -2x + 3y + 2z &= 8 \end{aligned}$$

Write down the augmented matrix and perform Gaussian elimination to bring the system to upper triangular form. Then, use back substitution to find the values of x, y, and z.

2. Consider the system of linear equations:

$$2x + 3y - z = 5$$

$$4x - y + 2z = 10$$

$$-2x + 2y - z = 3$$

Perform row reduction to bring the augmented matrix to echelon form. Write down the resulting system of equations and determine if the system is consistent or inconsistent.

3. Consider the following matrix:

$$\begin{bmatrix} 2 & 4 & 6 & 8 \\ 1 & 3 & 5 & 7 \\ 3 & 6 & 9 & 12 \end{bmatrix}$$

Perform row operations to reduce the matrix to its row-echelon form.

4. How is the augmented matrix used in determining the existence and uniqueness of solutions?

5. What information does the augmented matrix provide about the system of linear equations?

6. How is the determinant of a matrix related to the existence of solutions?

7. Can you have a consistent system with a zero determinant?

8. What does it mean for a system of linear equations to be consistent?

9. How can you determine if a system of linear equations is consistent?

10. Can a consistent system have a unique solution, infinitely many solutions, or no solution?

11. What is a homogeneous system of linear equations?

12. Can a homogeneous system have a non-trivial solution?

13. How does the trivial solution relate to homogeneous systems?

□□□

CHAPTER

7

SIMPLE LINEAR REGRESSION

Chapter Outline

• Introduction • Simple Linear Regression Model • Estimation of the Regression Parameters (β_0, β_1) • Meaning of Regression Parameters • Properties of the Least Squares Estimators • Coefficient of Determination

7.1 INTRODUCTION

A model provides a theoretical framework for better understanding of a phenomenon of interest. Thus, a model is a mathematical construct which is used to explain the mechanism that generated the observations at hand. In the biological, physical, and social sciences, as well as in business and engineering, linear models are useful in both the planning stages of research and analysis of the resulting data. The regression models are used for prediction, data description or explanation, parameter estimation, variable selection, and control of output. The scientific method is frequently used as a guided approach to learning. Linear statistical methods are widely used as a part of this learning process.

There are three types of linear models, that is

- Simple Linear Regression Model
- Multiple Linear Regression Model
- Analysis of Variance Model (ANOVA)

7.2 SIMPLE LINEAR REGRESSION MODEL

Regression analysis utilizes the statistical relation between two or more variables so that one variable can be predicted from the other variable(s).

In simple linear regression, we attempt to model the relationship between two variables, as seen in the following examples.

- Income and number of years of education,
- Height and weight of people,
- Length and width of envelopes,
- Temperature and output of an industrial process,
- Altitude and boiling point of water, or
- Dose of a drug and response.

In this section, we shall consider a regression model where there is only one independent variable and the regression function is linear in the regression parameters. The model can be stated as,

$$\boxed{y = \beta_0 + \beta_1 x + \varepsilon.}$$

where

(*i*) β_0 and β_1 are regression parameters. β_0 is the intercept and β_1 is the slope parameter.

(*ii*) x is the independent (or predictor) variable. We assume that it can be measured without error (i.e. x is not a random variable in the regression model).

(*iii*) ε is error term and $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$. The error term is not observable. In this context, error does not mean a mistake but is a statistical term representing random fluctuations, measurement errors, or the effect of factors outside our control.

(*iv*) y is the dependent (or response) variable.

The designation simple indicates that there is only one x to predict the response y, and linear means that the model is linear in β_0 and β_1.

The simple linear regression model for n observations can be written as:

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1 \quad \text{for all } i = 1, 2, 3, ..., n$$

In this case, y_i and ε_i are random variables and x_i are known constants.

Therefore, these assumptions hold for the above model:

(*a*) $E[\varepsilon_i] = 0$ for $i = 1, 2, 3, ..., n$, or equivalently $E[y_i] = \beta_0 + \beta_1 x_i$

(*b*) $Var[\varepsilon_i] = \sigma^2$ for $i = 1, 2, 3, ..., n$, or equivalently $Var[y_i] = \sigma^2$

(*c*) $Cov[\varepsilon_i, \varepsilon_j] = 0$ for $i \neq j$ or equivalently $Cov[y_i, y_j] = 0$.

Assumption (*a*) imply that y_i depends only on x_i and that all other variation in y_i is random.

Assumption (*b*) is also known as the assumption of homoscedasticity, homogeneous variance, or constant variance.

Assumption (*c*) implies the ε_i variables are uncorrelated.

7.3 ESTIMATION OF THE REGRESSION PARAMETERS (β_0, β_1)

In practical scenarios, the regression parameters are typically not known before hand. However, by utilizing a random sample of n observations represented as: (x_1, y_1), (x_1, y_1), ..., (x_n, y_n) we have the means to estimate the parameters β_0, β_1 and σ^2. These estimations, denoted as β_0 and β_1, are obtained through the method of least squares. Importantly, this method does not necessitate any assumptions regarding the distribution of the data.

In the method of least squares the point estimates of β_0 and β_1 are the values that minimize the sum of squared error (SSE) between observed values y and predicted values $\hat{y}$:

$$\boxed{\begin{aligned} SSE &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \end{aligned}}$$

Σ represents the sum over all data points ($i = 1$ to n).

y_i is the observed value for the i^{th} data point.

$\hat{y}_i$ is the predicted value for the i^{th} data point.

To find the minimum of SSE, we take partial derivatives with respect to β_0 and β_1 and set them equal to zero.

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum (y_i - \hat{y}_i)^2 = 0$$

$$\frac{\partial SSE}{\partial \beta_1} = -2\sum x_i (y_i - \hat{y}_i)^2 = 0$$

Solving for $\hat{\beta}_0$ and $\hat{\beta}_1$

Solving the above equations gives us the least squares estimates for β_0 and β_1.

For $\hat{\beta}_0$:

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

where:

$\bar{y}$ is the mean of the observed values (y)

$\bar{x}$ is the mean of the predictor variable (x).

For $\hat{\beta}_1$:

$$\boxed{\hat{\beta}_1 = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\Sigma_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\Sigma_{i=1}^{n} x_i^2 - n\bar{x}^2}}$$

These equations provide the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that define the best fitting linear regression line, minimizing the sum of squared differences between the observed and predicted values. These estimates are used to make predictions and describe the linear relationship between the predictor variable (x) and the response variable (y).

7.4 MEANING OF REGRESSION PARAMETERS

(*i*) In linear regression, $\hat{\beta}_1$: represents the slope or coefficient associated with the predictor variable x. The coefficient tells us how much the expected value (mean) of the dependent variable y is expected to change for a one-unit increase in x. In other words, it quantifies the impact of x on the predicted values of y. If $\hat{\beta}_1$ is positive, it indicates that an increase in x is associated with an increase in the expected value of y, and if $\hat{\beta}_1$ is negative, it indicates the opposite.

(*ii*) The intercept $\hat{\beta}_0$ represents the predicted value of the dependent variable y when the predictor variable x is equal to zero. In this context, it provides the starting point or baseline value of y when all other predictors are held constant. If x is a meaningful predictor variable that can take on a value of zero in the context of the problem, then the intercept $\hat{\beta}_0$ has a specific interpretation as: $\hat{\beta}_0 = E(y)$ at $x = 0$.

(*iii*) In cases where the predictor variable x does not have a meaningful interpretation at $x = 0$, the intercept $\hat{\beta}_0$ remains relevant as the baseline value of y when all other predictors are held constant. However, the coefficient $\hat{\beta}_1$ for x still retains its meaning as the change in the expected value of y associated with a one-unit change in x, even though $x = 0$ might not be a meaningful point within the context of the problem.

Example 7.1 Students in a statistics class (taught by one of the authors) claimed that doing the homework had not helped prepare them for the midterm exam. The exam score y and homework score x averaged up to time of the midterm) for the 1 students in the class were as follows:

x	5	6	7	8	9	10	11	12	13	14	15
y	1	4	3	8	7	7	13	10	16	17	13

x_i	y_i	x_i^2	$x_i y_i$
5	1	25	5
6	4	36	34
7	3	49	21
8	8	64	64
9	7	81	63
10	7	100	70
11	13	121	143]
12	10	144	120
13	16	169	208
14	17	196	238'
15	13	225	195

$$\sum x_i = 110$$

$$\sum y_i = 99$$

$$\sum x_i^2 = 1210$$

$$\sum x_i y_i = 1151$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{110}{11} = 10$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{99}{11} = 9$$

$$n\bar{x}\,\bar{y} = 11 \times 10 \times 9 = 990$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = (1151 - 990)/(1210 - 11 \times 10 \times 10)$$

$$= 1.4636$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 9 - 1.4636 \times 10 = -5.6364$$

The prediction equation is thus given by:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -5.6364 + 1.4636\,x$$

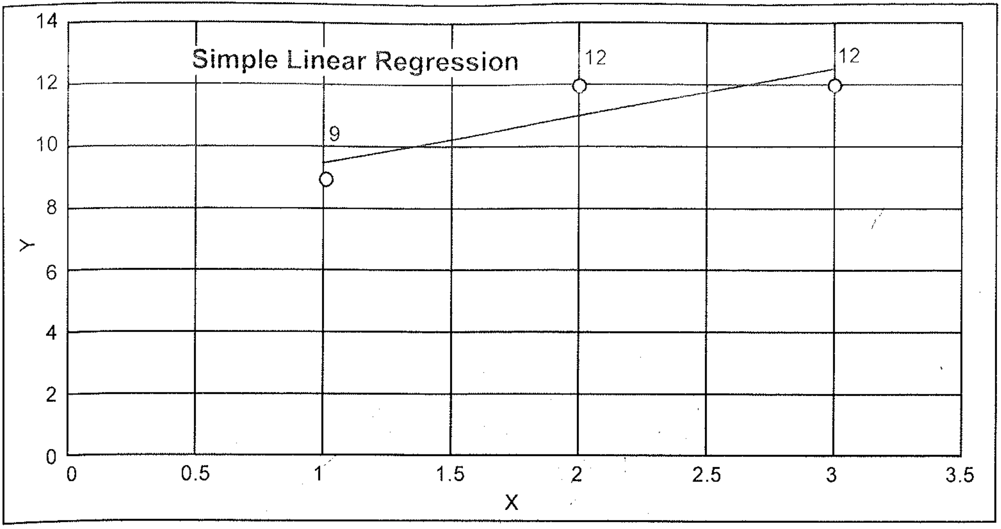


Fig. 7.1 Regression of *y* on *x*

7.5 PROPERTIES OF THE LEAST SQUARES ESTIMATORS

However, if the three assumptions hold, then

(a) The least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased i.e.,

$$E\left[\hat{\beta}_1\right] = \beta_1 \text{ and}$$

$$E\left[\hat{\beta}_0\right] = \beta_0$$

(b) And have minimum variance among all linear unbiased estimators (Gauss-Markov theorem)

Proof (a)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and using the assumption (a) $E[y_i] = \beta_0 + \beta_1 x_i$

$$E[\overline{y}] = \beta_0 + \beta_1 x$$

$$E[y_i - \overline{y}] = \beta_1 (x_i - \overline{x})$$

$$E\left[\hat{\beta}_1\right] = E\left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, E[(y_i - \bar{y})]}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, \beta_1 (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \beta_1$$

Therefore $\hat{\beta}_1$ is an unbiased estimator of β_1.

$$E\left[\hat{\beta}_0\right] = E\left[\bar{y} - \hat{\beta}_1 \bar{x}\right] = \beta_0 + \beta_1 \bar{x} - \bar{x}\, E\left[\hat{\beta}_1\right] = \beta_0$$

The estimator of $\hat{\beta}_0$ is also unbiased.

(b) Variance of Regression Parameter

(i) $$\text{Var}\left[\hat{\beta}_1\right] = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(ii) $$\text{Var}\left[\hat{\beta}_0\right] = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right\}$$

(i) Proof:

$\text{Var}[y_i] = \text{Var}[\beta_0 + \beta_1 x_i + \varepsilon_i] = \text{Var}[\varepsilon_i] = \sigma^2$ (because $\beta_0 + \beta_1 x_i$ is fixed).

$$\text{Var}\left[\hat{\beta}_1\right] = \text{Var}\left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]$$

Using the property of variance

$$\text{Var}[ax] = a^2\, \text{Var}[x]$$

Hence, we can write $\text{Var}\left[\hat{\beta}_1\right]$ as:

$$\text{Var}\left[\hat{\beta}_1\right] = \frac{1}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2}\, \text{Var}\left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right]$$

Apply Linearity of Expectation:

$$\text{Var}\left[\hat{\beta}_1\right] = \frac{1}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2} \sum_{i=1}^{n} \text{Var}[(x_i - \bar{x})(y_i - \bar{y})]$$

$$= \frac{1}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\,\mathrm{Var}\,[y_i-\bar{y}]$$

$$= \frac{1}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\,\mathrm{Var}\,[y_i]$$

$$\mathrm{Var}\left[\hat{\beta}_1\right] = \frac{1}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

Note that the variance of $\hat{\beta}_1$ is minimum when $\sum_{i=1}^{n}(x_i-\bar{x})^2$ is as large as possible and because of this reason, values of X should be chosen so as to cover the entire range of its values.

(ii) Proof:

$$\mathrm{Var}\left[\hat{\beta}_0\right] = \mathrm{Var}[\bar{y}-\hat{\beta}_1\bar{x}] = \mathrm{Var}[\bar{y}] + \bar{x}^2\,\mathrm{Var}\left[\hat{\beta}_1\right] - 2\,Cov\left[\bar{y},\hat{\beta}_1\right] = \beta_0$$

$$Cov\left[\bar{y},\hat{\beta}_1\right] = 0 \text{ as } \bar{y} \text{ and } \hat{\beta}_1 \text{ are uncorrelated.}$$

$$\mathrm{Var}\,[\bar{y}] = \frac{\sigma^2}{n}$$

$$\mathrm{Var}\left[\hat{\beta}_0\right] = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\mathrm{Var}\left[\hat{\beta}_0\right] = \sigma^2\left\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\};$$

The variance of $\hat{\beta}_0$ is minimum when $\bar{x} = 0$.

7.6 COEFFICIENT OF DETERMINATION

The coefficient of determination r^2 is defined as:

$$\boxed{r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{SSE}{SST}}$$

where $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ is the regression sum of squares and $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares. The total sum of squares can be partitioned into $SST = SSR + SSE$, that is,

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Thus r^2 gives the proportion of variation in y that is explained by the model or, equivalently, accounted for by regression on x_i.

x_i	y_i	x_1^2	$x_i y_i$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \bar{y})^2$
5	1	25	5	53.559	64
6	4	36	24	34.2787	25
7	3	49	21	19.2826	36
8	8	64	64	8.57084	1
9	7	81	63	2.1433	4
10	7	100	70	1.6E-07	4
11	13	121	143	2.14095	16
12	10	144	120	8.56616	1
13	16	169	208	19.2756	49
14	17	196	238	34.2693	64
15	13	225	195	53.5473	16

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 235.64$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = 280$$

$$r^2 = \frac{SSR}{SST} = \frac{235.64}{280} = 0.84159$$

The correlation between homework score and exam score is $r = \sqrt{0.8419} = 0.91736$.

PROBLEMS

1. What is the fundamental purpose of simple linear regression in machine learning?

2. How is the relationship between the dependent and independent variables represented in a simple linear regression equation?

3. What are the two main parameters estimated in simple linear regression, and how are they interpreted?

4. In the context of simple linear regression, what does the slope of the regression line signify?

5. How is the quality of fit assessed in simple linear regression?

6. What is the difference between the actual values and the predicted values in simple linear regression called?

7. How does the method of least squares contribute to finding the best-fitting line in simple linear regression?

8. Can simple linear regression be used for predictive modelling, and if so, how?

9. In simple linear regression, what is the role of the y-intercept in the regression equation?

10. What is the primary assumption underlying simple linear regression regarding the distribution of errors?

11. Consider the following data set that shows the relationship between the number of years of experience (independent variable, x) and the salary (dependent variable, y) of employees:

Years of Experience (x):	2	4	6	8	10
Salary (y) in $1000s:	50	60	65	72	80

(a) Calculate the mean of years of experience ($\overline{x}$) and the mean of salaries ($\overline{y}$).

(b) Determine the slope (m) and intercept (b) for the simple linear regression equation $y = mx + b$

(c) Write the equation of the regression line.

(d) Predict the salary for an employee with 5 years of experience.

12. Consider the following dataset representing the relationship between the number of hours spent on studying (independent variable, x) and the exam scores (dependent variable, y) for a group of students:

Hours (x):	4	6	8	10	12	14	16	18	20
Scores (y):	60	65	70	75	80	85	90	95	98

(*a*) Calculate the mean of hours ($\overline{x}$) and exam scores ($\overline{y}$).

(*b*) Determine the slope (m) and intercept (b) for the simple linear regression equation $y = mx + b$

(c) Write the equation of the regression line.

(*d*) Predict the exam score for a student who studies 15 hours.

13. You are provided with a dataset illustrating the relationship between the number of miles driven (x) and the fuel cost (y) for a larger sample of cars:

Miles Driven (x):	100	200	300	400	500	600	700	800	900
Fuel Cost (y) in $:	20	35	45	55	65	75	85	95	105

(*a*) Calculate the mean of miles driven ($\overline{x}$) and fuel cost ($\overline{y}$).

(*b*) Find the slope (m) and intercept (b) for the simple linear regression equation, $y = mx + b$

(*c*) Write the equation of the regression line.

(*d*) Predict the fuel cost for a car that has driven 250 miles.

❑❑❑

CHAPTER

8

MULTIPLE LINEAR REGRESSION MODEL

Chapter Outline

• Introduction • Multiple Linear Regression Model • Meaning of Regression Parameter • Least-Squares Estimator for β • Properties of the Estimated Parameters • Geometry of Least Squares

8.1 INTRODUCTION

The general linear model, often referred to as the multiple linear regression model, is used for analysing data involving multiple independent variables. In this context, we consider a linear multiple regression model with k independent variables. Multiple linear regression extends the principles of simple linear regression to accommodate situations where there are more than one independent variable influencing the dependent variable.

8.2 MULTIPLE LINEAR REGRESSION MODEL

The representation of the regression model is as follows:

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon}$$

Where,

(*i*) The regression parameters, denoted as $\beta_0, \beta_1, \dots, \beta_k$, represent the key coefficients within the model.

(*ii*) The independent variables, namely $x_1, x_2, \dots, x$, are the predictor variables. These variables are considered fixed and can be measured without any error or uncertainty.

(*iii*) The variable y, which serves as the response variable, is observable and subject to random error u.

This regression model comes into play when the response variable relies on multiple independent variables.

The regression function, also known as the response function, associated with the above model can be expressed as:

$$E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

The outcome y is frequently impacted by multiple predictor variables. For instance, the crop yield could be influenced by the quantities of nitrogen, potash, and phosphate fertilizers applied. While these factors can be manipulated by the experimenter, the yield might also be affected by uncontrollable variables, such as those related to weather conditions.

Let's consider that we have a random sample consisting of n observations of Y, collected at specific values of the independent variables.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

The assumptions for ε_i or y_i are the same as those for simple linear regression:

(a) $E[\varepsilon_i] = 0$ for $i = 1,2,3 \ldots, n$, or equivalently $E[y_i] = \beta_0 + \beta_1 x_{1i} \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$

(b) $Var\,[\varepsilon_i] = \sigma^2$ for $i = 1, 2, 3 \ldots, n$, or equivalently $Var\,[y_i] = \sigma^2$

(c) $Cov\,[\varepsilon_i, \varepsilon_j] = 0$ for $i \neq j$ or equivalently $Cov[y_i, y_j] = 0$

Assumption 1 states that the model is correct, in other words that all relevant x's are included, and the model is indeed linear. Assumption 2 asserts that the variance of y is constant and therefore does not depend on the x's. Assumption of 3 states that the y's are uncorrelated with each other, usually holds in a random sample (the observations would typically be correlated in a time series or when repeated measurements are made on a single plant or animal). Later, we will add a normality assumption, under which the y variable will be independent as well as uncorrelated.

When all three assumptions hold, the least-squares estimators of the β's have some good properties. If one or more assumptions do not hold, the estimators may be poor. Under the normality assumption, the maximum likelihood estimators have excellent properties.

Any of the three assumptions may fail to hold with real data. Several procedures have been devised for checking the assumptions. These diagnostic techniques are discussed in Chapter 10.

We get the following set of n equations;

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \ldots + \beta_k x_{k1} + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{12} \beta_2 x_{22} + \ldots + \beta_k x_{k2} + \varepsilon_2$$
$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots + \beta_k x_{kn} + \varepsilon_n$$

These n equations can be written in matrix form as:

$$\boxed{Y = X\beta + \varepsilon} \qquad \ldots (8.1)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \ldots(8.2)$$

where,

(i) Y represents a column vector consisting of n observed values.

(ii) X is an (nxp) matrix comprising known values, where $p = (k + 1)$.

(iii) β is a column vector representing p unknown parameters.

(iv) ε is a column vector that contains unobservable errors associated with n observations, and it's important to note that p is less than n.

We assume that the Rank (X) = p. In such a scenario, the matrix X is considered to have full rank.

The preceding three assumptions on ε_i or y_i can be expressed in terms of the model

(a) $E[\varepsilon] = 0$ or $E[Y] = X\beta$.

(b) $Cov(\varepsilon) = \sigma^2 I$ or $Cov(y) = \sigma^2 I$

8.3 MEANING OF REGRESSION PARAMETER

(i) The parameter β_i signifies the alteration in $E(y)$ for each unit increase in x_i while keeping all other variables constant.

(ii) If the model encompasses $x_1 = 0, x_2 = 0, \ldots, x_k = 0$, the intercept β_0 represents the expected value $E(y)$ when all independent variables are set to zero.

When the influence of x_i on $E(y)$ remains consistent regardless of the values of the other variables for all i, it can be said that the variables exhibit an additive effect or do not interact with each other.

8.4 LEAST-SQUARES ESTIMATOR FOR β

The least squares method is employed to calculate estimates for the p unknown regression parameters. It's important to note that no distributional assumptions regarding y are necessary to obtain these estimators.

For the parameters $\beta_0, \beta_1, \ldots, \beta_k$, our objective is to find estimators, denoted as $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, that minimize the sum of squared deviations between the n observed values of y and their corresponding predicted values $\hat{y}$. These estimations are the values that minimize the sum of squared errors (SSE) between observed values y and predicted values $\hat{y}$.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SSE = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_k x_{ki})^2 \qquad \ldots(8.6)$$

Note that the predicted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}$ estimates $E(y_i)$, not $\hat{y}_i$. A better notation would be $E[\hat{y}_i]$ but $\hat{y}_i$ is commonly used.

To find the values of $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k$, that minimize SSE, we could differentiate SSE with respect to each β_1 and set the results equal to zero to yield $p = k + 1$ equations that can be solved simultaneously for the $\hat{\beta}_j$'s.

$$\frac{\partial SSE}{\partial \beta_1} = 0$$

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum (y_i - \hat{y}_i)^2 = 0 \qquad \ldots(1)$$

$$\frac{\partial SSE}{\partial \beta_1} = -2\sum x_{1i}(y_i - \hat{y}_i)^2 = 0 \qquad \ldots(2)$$

$$\vdots$$

$$\frac{\partial SSB}{\partial \beta_k} = -2\sum x_{ki}(y_i - \hat{y}_i)^2 = 0$$

By solving above equations, we get following p set of equations:

$$\Sigma \hat{y}_i = n\hat{\beta}_0 + \hat{\beta}_1 \Sigma x_{1i} + \hat{\beta}_2 \Sigma x_{2i} + \ldots + \hat{\beta}_k \Sigma x_{ki}$$

$$\Sigma x_{1i}\, \hat{y}_i = \hat{\beta}_0 \Sigma x_{1i} + \hat{\beta}_1 \Sigma x_{1i}^2 + \hat{\beta}_2 \Sigma x_{1i} x_{2i} + \ldots + \hat{\beta}_k \Sigma x_{1i} x_{ki}$$

$$\ldots\ldots$$

$$\Sigma x_{ki}\, \hat{y}_i = \hat{\beta}_0 \Sigma x_{ki} + \hat{\beta}_1 \Sigma x_{ki} x_{1i} + \hat{\beta}_2 \Sigma x_{ki} x_{2i} + \ldots + \hat{\beta}_k \Sigma x_{ki}^2$$

The equations can be written in a matrix form as,

$$X^T X \hat{\beta} = X^T y$$

Pre-multiply both sides by, $(X^T X)^{-1}$ to get,

$$(X^T X)^{-1} X^T X \hat{\beta} = (X^T X)^{-1} X^T y$$

The least squares estimate of the parameters is given by,

$$\boxed{\hat{\beta} = (X^T X)^{-1} X^T y}$$

Because we assumed that X is of full rank, the matrix $(X^T X)$ has an inverse, and the solution of $\hat{\beta}$ is unique.

Example 8.1 Consider a data set,

Area (x_1)	**Rooms (x_2)**	**Age (x_3)**	**Price (y)**
23	3	8	6562
15	2	7	4569
24	4	9	6897
29	5	4	7562
31	7	6	8234
25	3	10	7485

Solution: Let us consider,

Here **area, rooms, age** are features/independent variables and **price** is the target/dependent variable.

$$
\begin{aligned}
X = {} & [[1, 23, 3, 8], \\
& [1, 15, 2, 7], \\
& [1, 24, 4, 9], \\
& [1, 29, 5, 4], \\
& [1, 31, 7, 6], \\
& [1, 25, 3, 10]] \\
X^T = {} & [[1, 1, 1, 1, 1, 1], \\
& [23, 15, 24, 29, 31, 25], \\
& [3, 2, 4, 5, 7, 3], \\
& [8, 7, 9, 4, 6, 10]] \\
y = {} & [6562, 4569, 6897, 7562, 8234, 7485]
\end{aligned}
$$

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 305.33 \\ 236.86 \\ -4.76 \\ 102.90 \end{bmatrix}$$

Hence,

$$\hat{\beta}_0 = 305.33$$

$$\hat{\beta}_1 = 236.86$$

$$\hat{\beta}_2 = -4.76$$

$$\hat{\beta}_3 = 102.90$$

The model can be interpreted as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$$

$$\hat{y}_i = 305.33 + 236.86\, x_{1i} + -4.76\, x_{2i} + 102.90\, x_{3i}$$

Actual value of Test Data = [8234, 7485]

Predicted value of Test Data = [8232. 7241.52380952]

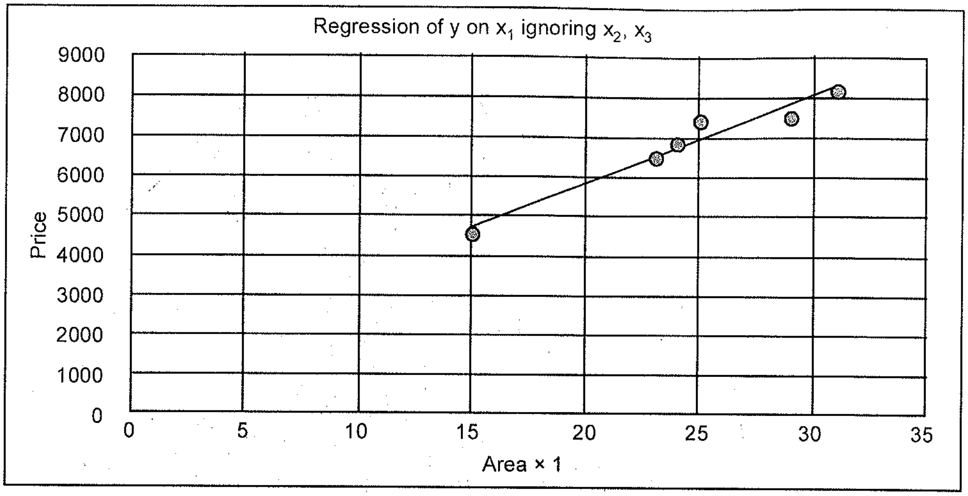


Fig. 8.1 Regression of y on x_1 ignoring x_2, x_3.

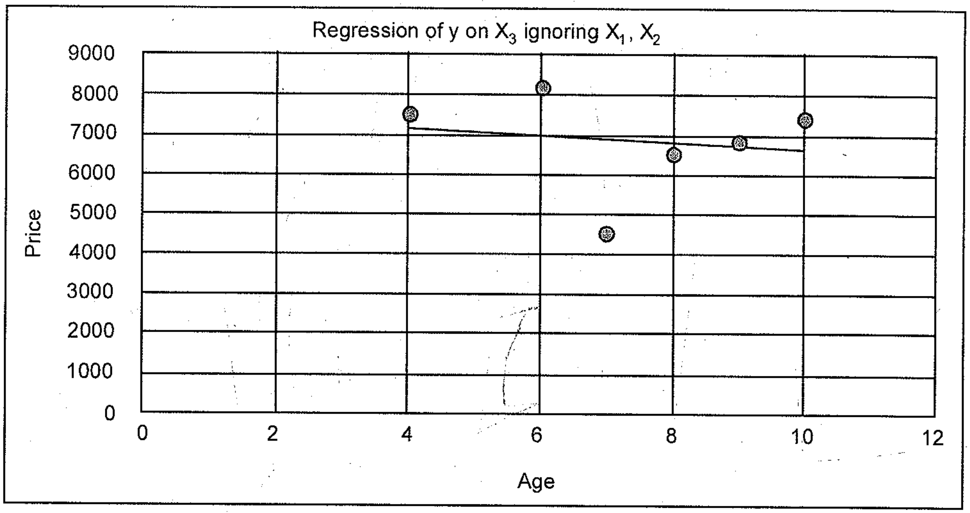


Fig. 8.2 Regression of y on x_3 ignoring x_1, x_2.

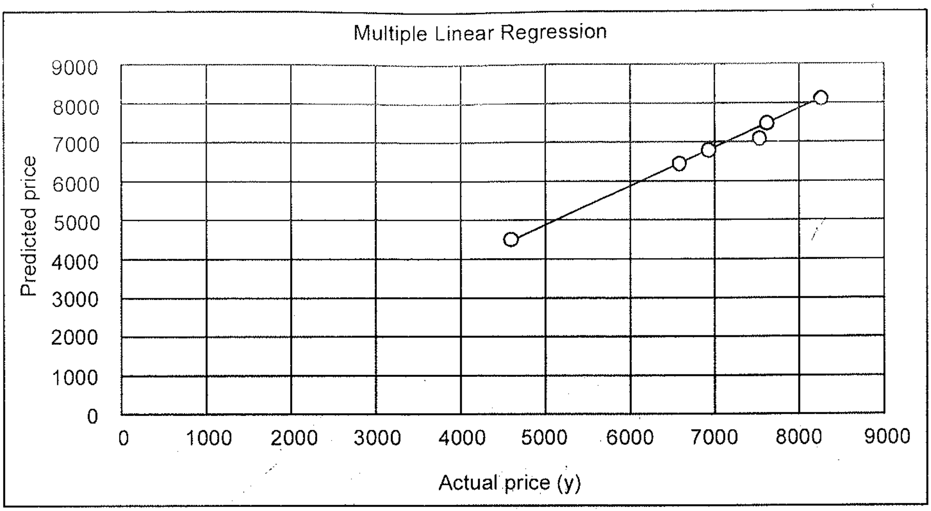


Fig. 8.3 Regression of y on x_3 ignoring x_1, x_2

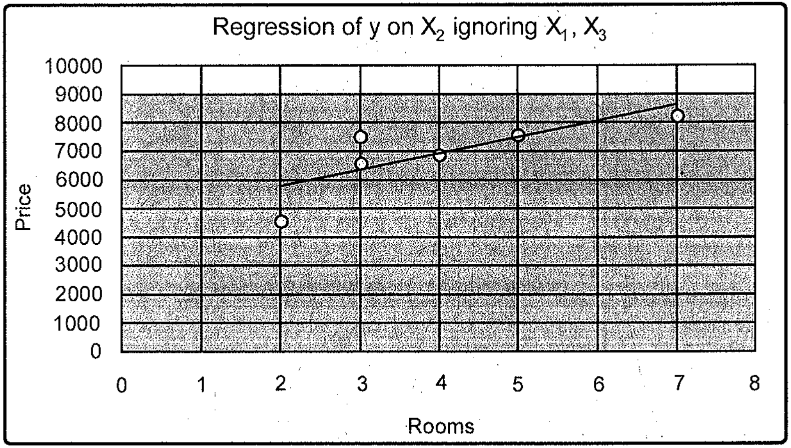


Fig. 8.4 Multiple Linear Regression

8.5 PROPERTIES OF THE ESTIMATED PARAMETERS

(a) $\hat{\beta}$ is unbiased.

(b) $Covar\left(\hat{\beta}\right) = (X^T X)^{-1} \sigma^2$

(c) $\hat{\beta}$ is a normal random vector.

(d) The least squares estimators have the minimum variance (are the most efficient) among the class of all the unbiased linear unbiased estimators of $\hat{\beta}$.

8.5.1 Unbiases of $(\hat{\beta})$

$$E[Y] = E[X\beta + \varepsilon] = X\beta$$

$$E[\hat{\beta}] = E[(X^T X)^{-1} X^T y] = (X^T X)^{-1} X^T X\beta = \beta$$

Therefore, $\hat{\beta}$ is unbiased. For each j, $\hat{\beta}_j$ is an unbiased estimator of $\hat{\beta}_j$.

8.5.2 Covar $(\hat{\beta}) = (X^T X)^{-1} \sigma^2$

We use the property of Covariance of a random vector.

Let V be a random vector. For a given matrix M of constant elements,

$cov(MV) = M\, cov(V)\, M^T$ Let $M = (X^T X)^{-1} X^T$ and $V = Y$

$$cov(\hat{\beta}) = (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1}$$

$$cov(\hat{\beta}) = (X^T X)^{-1} \sigma^{-2}$$

The i^{th} diagonal element of the matrix $cov(\hat{\beta}) = (X^T X)^{-1} \sigma^2$ gives the variance of $\hat{\beta}_i$ and the $(i, j)^{th}$ element gives the covariance of $\hat{\beta}_i$ and $\hat{\beta}_j$ for $i \neq j$.

The variance matrix $cov(\hat{\beta})$ is a (pxp) symmetric matrix. Because it gives the variance as well as the covariance, it is also referred to as the Variance-Covariance matrix of $\hat{\beta}$. The matrix is a positive definite matrix if and only if all the eigenvalues are greater than zero

$\hat{\beta} = (X^T X)^{-1} X^T y$ is a linear function of normal vector y. Therefore $\hat{\beta}$ is also a normal vector.

Example 8.2

(*a*) Use the matrix method to estimate the parameters of the simple linear regression model.

(*b*) Obtain the variance-covariance matrix of the estimators.

Solution.

(*a*) We have $\quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, ..., n$

Using Matrix notation, the set of n equation can be written as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y = X\beta + \varepsilon$$

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_1 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$X^T X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$(X^T X)^{-1} = \left[\frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}\right] \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T X)^{-1} X^T y$$

$$= \left[\frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}\right] \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$= \left[\frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}\right] \begin{bmatrix} \Sigma x_i^2 \Sigma y_i - \Sigma x_i \Sigma x_i y_i \\ -\Sigma x_i \Sigma y_i + n\Sigma x_i y_i \end{bmatrix}$$

We obtain,

$$\hat{\beta}_0 = \frac{\bar{y}\Sigma x_i^2 - \bar{x}\,\Sigma x_i y_i}{\Sigma x_i^2 - n\bar{x}^2}$$

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i - n\bar{y}\,\bar{x}}{\Sigma x_i^2 - n\bar{x}^2} \quad \text{or}$$

(b) $$\text{Var}(\hat{\beta}) = \text{Var}((X^TX)^{-1}X^Ty) = (X^TX)^{-1}\sigma^2.$$

$$(X^TX)^{-1} = \left[\frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}\right]\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

$$\left[\frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}\right]\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}\sigma^2 = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0,\hat{\beta}_1) \\ Cov(\hat{\beta}_0,\hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix}$$

$$\text{Var}(\hat{\beta}_0) = \sigma^2\frac{\Sigma x_i^2}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{\sigma^2}{n} + \overline{x}^2\frac{\sigma^2}{\Sigma(x_i - \overline{x})^2}$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma(x_i - \overline{x})^2}$$

$$\text{Cov}(\hat{\beta}_0,\hat{\beta}_1) = \frac{-\sigma^2\,\Sigma x_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{-\overline{x}\,\sigma^2}{\Sigma(x_i - \overline{x})^2}$$

8.5.3 Gauss-Markov Theorem

If $E[y] = X\beta$ and $cov(y) = \sigma^2 I$, the least-squares estimators $\hat{\beta}_j$ for all $j = 0, 1, 2, \ldots, k$ have minimum variance among all linear unbiased estimators.

Proof: We examine a linear estimator Ay for β and aim to determine the matrix A that results in Ay being a minimum variance unbiased estimator of β. For Ay to qualify as an unbiased estimator of β. It's necessary that:

$$E(Ay) = \beta$$

Using the assumption $E(y) = \beta$, this can be expressed as:

$$E(Ay) = AE(y) = AX\beta = \beta,$$

Which gives the unbiased condition $AX = I$

Since the relationship $AX\beta = \beta$ must hold for any possible value of β.

The covariance matrix for the estimator Ay is given by

$$cov(Ay) = A(\sigma^2 I)A^T = \sigma^2 AA^T$$

The variances of the $\hat{\beta}_j$"s are specifically located along the diagonal of the $\sigma^2 AA^T$ matrix. Consequently, our goal is to select matrix A, while adhering to the constraint $AX = I$, in such a way that minimizes the diagonal elements of AA^T.

To relate Ay and $\hat{\beta} = (X^TX)^{-1}X^TY$, we apply an operation of addition and subtraction involving $(X^TX)^{-1}X^T$, resulting in the following expression:

$$AA^T = (A - (X^TX)^{-1}X^T + (X^TX)^{-1}X^T)\,(A - (X^TX)^{-1}X^T + (X^TX)^{-1}X^T)^T$$

Expanding this expression with respect to $A - (X^T X)^{-1} X^T$ and $(X^T X)^{-1} X^T$, we obtain four terms. Notably, two of these terms become zero due to the constraint $AX = I$. Consequently, the result is:

$$AA^T = (A - (X^T X)^{-1} X^T)\,(A - (X^T X)^{-1} X^T)^T + (X^T X)^{-1}$$

The matrix $(A - (X^T X)^{-1} X^T)\,(A - (X^T X)^{-1} X^T)^T$ is positive semidefinite and the diagonal elements are greater than or equal to zero. These diagonal elements can be made equal to zero by choosing $A = ((X^T X)^{-1} X^T$. This value of A also satisfies the unbiasedness condition $AX = I$. The resulting minimum variance estimator of β is $Ay = (X^T X)^{-1} X^T y$ which is equal to the least-squares estimator of $\hat{\beta}$.

The Gauss-Markov theorem is sometimes stated as follows: When we have $E[y] = X\beta$ and $cov(y) = \sigma^2 I$, the least-squares estimators $\hat{\beta}_j$ for all $j = 0, 1, 2, ..., k$ are regarded as the Best Linear Unbiased Estimators (BLUE). In this context, 'best' implies they have the minimum variance, and 'linear' indicates that these estimators are linear functions of y.

What makes the Gauss-Markov theorem remarkable is its wide-ranging applicability. This theorem remains valid for any distribution of y; it does not necessitate the assumption of normality. The sole assumptions utilized in the proof are $E[y] = X\beta$ and $cov(y) = \sigma^2 I$. If these assumptions are not satisfied, β might be biased, or each $\hat{\beta}_j$ could have a higher variance compared to some other estimator.

8.6 GEOMETRY OF LEAST SQUARES

In Sections 8.1 – 8.5, we introduced the multiple linear regression model as the matrix equation $Y = X\beta + \varepsilon$. We established the concept of least-squares estimation by considering deviations from the model. Subsequently, we employed matrix calculus and algebra to derive the estimators for β.

Now, we will provide an alternative but equally valid derivation of these estimators, entirely based on geometric concepts.

In two dimensions, we illustrated the least squares principle by creating a scatter plot (as seen in Fig. 7.1) of n data points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. We then visualized the least-squares regression line as the optimal straight line that fits the data. While this approach can be extended to represent the least-squares estimate in multiple linear regression by finding the best-fitting hyperplane in $(k + 1)$-dimensional space fornpoints $(x_{11}, x_{12}, ..., x_{1k}, y_1)$, $(x_{21}, x_{22}, ..., x_{2k}, y_2)$, ..., $(x_{n1}, x_{n2}, ..., x_{nk}, y_n)$, it's important to note that the geometric approach to least-squares estimation in multiple linear regression does not necessitate this high-dimensional generalization.

The forthcoming geometric approach holds a distinct appeal due to its mathematical elegance. Notably, it allows us to derive estimators without the need for matrix calculus. Furthermore, this geometric approach offers a deeper and more insightful understanding of statistical inference.

It's worth noting that several advanced statistical techniques, such as kernel smoothing (Eubank and Eubank 1999), Fourier analysis (Bloomfield 2000), and wavelet analysis (Ogden 1997), can be seen as extensions or generalizations of this geometric approach.

The origins of the geometric approach to linear models can be traced back to Fisher (Mahalanobis 1964). Additionally, Christensen (1996) and Jammalamadaka and Sengupta (2003) have extensively explored the linear statistical model from a predominantly geometric perspective.

8.6.1 Subspace Formulation of Linear Models

The geometric approach to least squares commences by considering two high-dimensional spaces: a $(k + 1)$dimensional space and an n-dimensional space. In this context, the unknown parameter vector β can be visualized as a singular point within the $(k + 1)$ dimensional space. Here, each axis corresponds to one of the $k + 1$ regression coefficients, namely $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Consequently, we refer to this space as the parameter space (as illustrated in Fig. 8.5).

Similarly, the data vector y can be envisioned as a singular point situated in an n-dimensional space, where each axis corresponds to one of the n observations. Accordingly, this space is referred to as the data space.

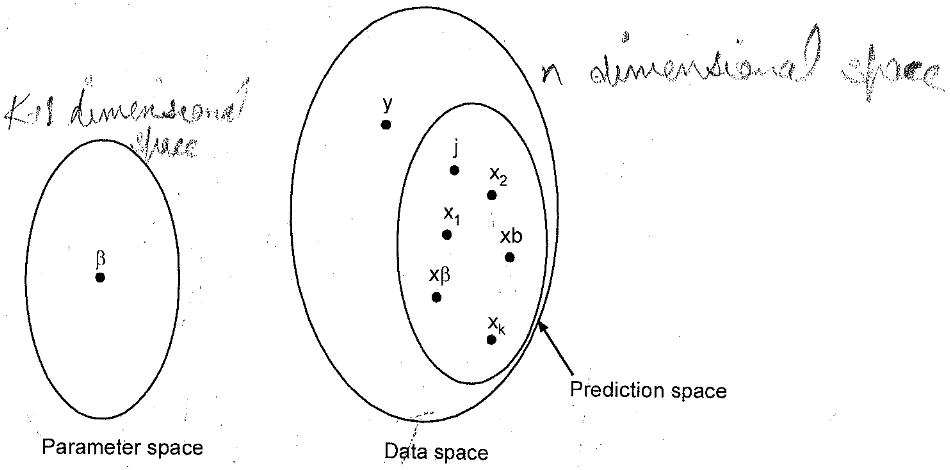


Fig. 8.5 Parameter space, data space, and prediction space with representative elements

The X matrix from the multiple regression model (8.4) can be expressed as a partitioned matrix, considering its $k + 1$ columns as:

$$X = (j, x_1, x_2, x_3, \ldots, x_k)$$

The columns of X, including column j, represent n-dimensional vectors, making them individual points within the data space. It's important to highlight that due to our assumption that X has a rank of $k + 1$, these vectors are linearly independent.

The set of all possible linear combinations of X's columns, forms a subset within the data space. Elements within this subset can be expressed as:

$$X = b_0 j + b_1 x_1 + b_2 x_2 + b_3 x_3, \ldots, b_k x_k$$

Here, b represents any vector of dimension $k + 1$, effectively encompassing all vectors within the parameter space. This specific subset possesses the characteristics of a subspace since it maintains closure under both addition and scalar multiplication (as discussed in Harville 1997, pp. 28-29).

This subset is commonly referred to as the subspace generated or spanned by the columns of X. In our discussions, we'll refer to this subspace as the prediction space. It's worth noting that the columns of X serve as the basis set for this prediction space.

8.6.2 Orthogonal Projections

The multiple linear regression model (8.4) posits that y can be represented as a vector within the prediction space, $E[y] = X\beta$, alongside a vector ε denoting random errors (as depicted in Fig. 8.6). The challenge lies in the fact that neither β nor ε is known. However, we do possess knowledge of the data vector y, which is not situated within the prediction space. It is also established that $E[y]$ resides within the prediction space.

To grasp multiple linear regression geometrically, consider it as a process where we aim to find a reasonable estimation of $E[y]$ within the prediction space. Subsequently, we determine the vector in the parameter space associated with this estimate (as illustrated in Fig. 8.6). This estimation of $E[y]$ is represented as $\hat{y}$ and the corresponding vector within the parameter space is denoted as $\hat{\beta}$.

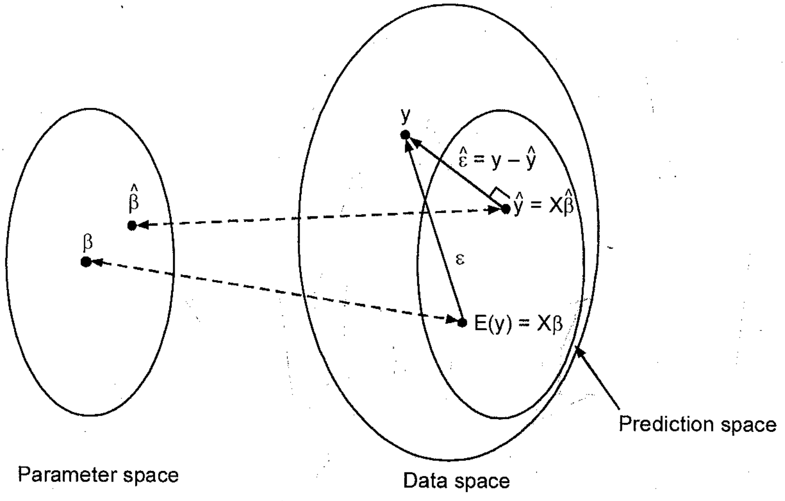


Fig. 8.6 Geometric relationships of vectors associated with the multiple linear regression model

A sensible geometric approach is to estimate $E[y]$ by identifying the point within the prediction space that is closest to y. Remarkably, it can be determined that $\hat{y}$, this closest point in the prediction space to y, must satisfy the condition that the difference vector $\hat{\varepsilon} = y - \hat{y}$ is orthogonal (perpendicular) to the prediction space (as outlined in Harville 1997,

p. 170). Additionally, since the prediction space is spanned by the columns of X, the point $\hat{y}$ must be such that $\hat{\varepsilon}$ is orthogonal to these columns of X. Consequently, we are led to seek $\hat{y}$ in a manner that satisfies the equation:

$$X^T\hat{\varepsilon} = 0$$

$$X^T(y-\hat{y}) = X^T(y - X\hat{\beta}) = X^Ty - X^TX\hat{\beta} = 0$$

which leads to,

$$X^TX\hat{\beta} = X^Ty$$

By relying solely on geometric concepts, we derive the normal equations and, as a result, obtain the conventional least-squares estimator $\hat{\beta}$ as presented in (8.4).

PROBLEMS

1. What is the difference between simple and multiple linear regression?

2. Explain the concept of multiple linear regression in simple terms.

3. What is the role of the intercept term in a multiple linear regression equation?

4. How is the residual in multiple linear regression calculated?

5. What does the term "adjusted R-squared" address in the context of multiple linear regression?

6. Explain the concept of multicollinearity and its potential impact on multiple linear regression.

7. In the context of multiple linear regression, what is the purpose of residual analysis?

8. Why is it important to standardize variables in multiple linear regression?

9. What is the difference between correlation and multicollinearity in the context of multiple linear regression?

10. Explain the concept of heteroscedasticity and its implications for multiple linear regression.

11. Consider a multiple linear regression model with two independent variables, X_1 and X_2, and the following coefficients: $\beta_0 = 10$, $\beta_1 = 3$, $\beta_2 = 4$. If $X_1 = 2$ and $X_2 = 5$, calculate the predicted value $\hat{Y}$.

12. Consider the following data for a multiple linear regression model with twoindependent variables X_1 and X_2 and a dependent variable Y:

X_1	X_2	Y
2	4	20
3	5	25
4	6	30
5	7	35

(*a*) Fit a multiple linear regression model and write down the estimated regression equation.

(*b*) Calculate the sum of squared residuals (SSE) for the model.

(*c*) Calculate the coefficient of determination R^2 for the model.

13. For a given multiple linear regression model, the coefficients are as follows:

$$\hat{Y} = 15 + 2X_1 - 3X_2$$

(*a*) Interpret the meaning of the intercept (15), the coefficient for X_1 (2), and the coefficient for X_2 (–3).

(*b*) If $X_1 = 4$ and $X_2 = 5$, calculate the predicted value Y?.

14. Suppose you have a dataset with three independent variables X_1, X_2, and X_3 and a dependent variable Y:

X_1	X_2	X_3	Y
3	4	2	25
5	6	3	35
7	8	4	45

(a) Fit a multiple linear regression model using this data and write down the estimated regression equation.

(b) Calculate the coefficient of determination R^2 for the model.

❑❑❑

CHAPTER

9

ANALYSIS OF VARIANCE (ANOVA) MODELS

Chapter Outline

- Introduction • Single Factor ANOVA Model • Analysis of Single Factor ANOVA • Two-Factor ANOVA Model • Factorial Experiment • The two-way Factorial Analysis • Analysis of Covariance

9.1 INTRODUCTION

ANOVA models are employed to investigate how specific factors impact the response variable Y, with factors encompassing both qualitative and quantitative independent variables. In a single-factor ANOVA model, the focus is solely on one particular factor, while in multifactor studies, two or more factors are concurrently assessed. Each factor consists of two or more levels, with these levels being termed treatments in the context of a single-factor model.

For instance, take monthly family income as an example. It can be categorized into three levels: (*i*) low, (*ii*) average, and (*iii*) high, constituting three income treatments. Income itself is a quantitative factor.

Similarly, consider the gender factor, which encompasses two levels or treatments: (*i*) male and (*ii*) female. Gender is a qualitative factor.

Analysis of Variance (ANOVA) possesses the following characteristics:

- It conducts a statistical assessment of variance among two or more sets of data.
- It assesses whether disparities in the means of various datasets arise from random variation or have a significant underlying cause.
- It employs the statistical F-ratio test.

There are two primary categories of variance analysis:

- Single Factor ANOVA model
- Two Factor ANOVA model

9.2 SINGLE FACTOR ANOVA MODEL

Consider a scenario in which a researcher has devised two chemical additives designed to enhance the fuel efficiency of gasoline. To create a model for this situation, we can begin with the premise that in the absence of additives, a gallon of gasoline typically provides an average of μ miles per gallon. Subsequently, if we introduce chemical additive 1, it is anticipated to boost mileage by t_1 miles per gallon, while the introduction of chemical additive 2 is expected to result in an increase of t_2 miles per gallon.

The model could be expressed as follows:

$$y_1 = \mu + t_1 + \varepsilon_1$$

$$y_2 = \mu + t_2 + \varepsilon_2 \qquad \text{...(9.1)}$$

Where y_1 is the miles per gallon from a tank of gasoline containing chemical 1 and ε_1 is the random error term distributed as $N(0, \sigma^2)$. Similarly, the variables y_2 and ε_2 are defined.

The researcher would like to estimate the parameters μ, t_1, and t_2 and test hypotheses such as $H_0 : t_1 = t_2$

To obtain reliable estimates, the researcher must collect mileage per gallon data for multiple refills of gasoline for each chemical additive. Let's imagine that the experiment involves refilling the tanks of six identical cars with gasoline. Three of these tanks receive chemical additive 1, while the remaining three tanks are treated with chemical additive 2. We can formulate a separate model for each of the six observations as follows:

$$y_{11} = \mu + t_1 + \varepsilon_{11}\ y_{12} = \mu + t_1 + \varepsilon_{12}\ y_{13} = \mu + t_1 + \varepsilon_{13} \qquad \text{...(9.2)}$$

$$y_{21} = \mu + t_2 + \varepsilon_{21}\ y_{22} = \mu + t_2 + \varepsilon_{22}\ y_{23} = \mu + t_2 + \varepsilon_{23}$$

Or $$y_{ij} = \mu + t_i + \varepsilon_{ij} \qquad \text{...(9.3)}$$

Where y_{ij} is the observed as miles per gallon of the j^{th} car that contains the i^{th} chemical in its tank and ε_{ij} is the associated random error. The six equations in (9.2) can be written in matrix form as:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ t_1 \\ t_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix} \qquad \text{...(9.4)}$$

Where $$y = X\beta + \varepsilon$$

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \mu \\ t_1 \\ t_2 \end{bmatrix}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix}$$

X is a 6×3 matrix whose rank is 2 since the first column is the sum of the second and third columns, which are linearly independent. Since X is not of full rank, we cannot directly estimate $\beta = (\mu, t_1, t_2)'$ and test hypotheses. Thus, for example, the parameters μ, t_1, and t_2 cannot be estimated by $\hat{\beta} = (X^T X)^{-1} X^T y$, because $(X^T X)^{-1}$ does not exist.

To delve deeper into understanding why the above model did not achieve full-rank status as a model, let's revisit the interpretation of its parameters. Initially, μ was introduced as the pre-chemical mean, while t_1 and t_2 were meant to signify the increments caused by chemicals 1 and 2, respectively. Nevertheless, it's important to note that the model $y_{ij} = \mu + t_i + \varepsilon_{ij}$ cannot uniquely support this characterization. For example, $\mu = 15, t_1 = 1$ and $t_2 = 3$, the model becomes:

$$y_{1j} = \mu + t_1 + \varepsilon_{1j} = 15 + 1 + \varepsilon_{1j} = 16 + \varepsilon_{1j}, \quad j = 1, 2, 3$$
$$y_{2j} = \mu + t_2 + \varepsilon_{2j} = 15 + 3 + \varepsilon_{2j} = 18 + \varepsilon_{2j}, \quad j = 1, 2, 3$$

However, from $y_{1j} = 16 + \varepsilon_{1j}$ and $y_{2j} = 18 + \varepsilon_{2j}$ we cannot determine that $\mu = 15$, $t_1 = 1$, and $t_2 = 3$, because the model can also be written as

$$y_{1j} = 10 + 6 + \varepsilon_{1j}, \qquad j = 1, 2, 3$$
$$y_{2j} = 10 + 8 + \varepsilon_{2j}, \qquad j = 1, 2, 3$$

or, alternatively, as

$$y_{1j} = 25 - 9 + \varepsilon_{1j}, \qquad j = 1, 2, 3$$
$$y_{2j} = 25 - 7 + \varepsilon_{2j}, \qquad j = 1, 2, 3$$

or in infinite other ways.

The parameters μ, t_1 and t_2 are not unique and therefore cannot be estimated. With three parameters and rank $(X) = 2$, the model is said to be over-parameterized. Note that increasing the number of observations (replications) for each of the two additives will not change the rank of X.

There are various ways each with its own advantages and disadvantages-to remedy this lack of uniqueness of the parameters in the over-parameterized model.

Three such approaches are:

1. Redefine the model using a smaller number of new parameters that are unique,
2. Use the over-parameterized model but place constraints on the parameters so that they become unique, and
3. In the over-parameterized model, work with linear combinations of the parameters that are unique and can be unambiguously estimated.

We briefly illustrate these three techniques.

(1) In order to reduce the parameter count, let's examine the above example again. In this, we have $y_{1j} = 16 + \varepsilon_{1j}$ and $y_{2j} = 18 + \varepsilon_{2j}$. The values 16 and 18 represent the means after applying two treatments. In a more general context, we can denote these means as μ_1 and μ_2, and the model can be expressed as follows:

$$y_{1j} = \mu_1 + \varepsilon_{1j} \quad \text{and} \quad y_{2j} = \mu_2 + \varepsilon_{2j}$$

The means μ_1 and μ_2 are unique and can be estimated. The redefined model for all six observations takes the form;

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix}$$

which we write as

$$Y = W\mu + \varepsilon$$

The matrix W is full-rank, and we can estimate μ as

$$\varepsilon = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix} \left(W^T W\right)^{-1} W^T y$$

This solution is called re-parameterization.

(2) An alternative to reducing the number of parameters is to incorporate constraints on the parameters μt_1, and t_2. We denote the constrained parameters as μ^*, t_1^*, and t_2^*. Considering the restriction $t_1^* + t_2^* = 0$ serves the distinct purpose of establishing μ^* as the new mean after the application of treatments, with t_1^* and t_2^* representing deviations from

this mean. With this constraint, $y_{1j} = 16 + \varepsilon_{1j}$ and $y_{2j} = 18 + \varepsilon_{2j}$, the model can be written only as:

$$y_{1j} = 17 - 1 + \varepsilon_{1j}, \quad y_{2j} = 17 + 1 + \varepsilon_{2j}.$$

This model is now unique because there is no other way to express it so that $t_1^* + t_2^* = 0$. Such constraints are often called side conditions. The model, $y_{ij} = \mu^* + t_i^* + \varepsilon_{ij}$ is subject to $t_1^* + t_2^* = 0$, which can be expressed in a full-rank format by using $t_2^* = t_1^* 1$ to obtain $y_{1j} = \varepsilon^* + t_1^* + \varepsilon_{1j}$ and $y_{2j} = \mu^* - t_1^* + \varepsilon_{2jj}$. The six observations can then be written in matrix form as

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \mu^* \\ t_1^* \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix}$$

Which we write as

$$y = X^* \mu^* + \varepsilon$$

The matrix X^* is full-rank, and the parameters μ^* and t_1^* can be estimated. It must be kept in mind, however, that specific constraints impose specific definitions on the parameters.

9.3 ANALYSIS OF SINGLE FACTOR ANOVA

In a single-factor ANOVA model, data points are categorized based on a single factor. For instance, when examining the factors affecting crop yield, such as soil quality, rainfall availability, seed quantity, and fertilizer usage, if the focus is solely on one of these factors, it constitutes a single-factor ANOVA model.

While investigating the impact of fertilizer on crop yield, the approach involves applying various types of fertilizers to distinct paddy fields and discerning the variations in the fertilizer's effects on yield.

We consider the ANOVA model where there is only one factor with r levels. For the i^{th} factor level, we have available a random sample of size n_i $i = 1, 2, .., r$.

We say there are n_i replications for the i^{th} treatment.

Let y_{ij} be the j^{th} observation of the i^{th} treatment $\forall\ i = 1, 2, \ldots r, j = 1, 2, .., n_i$

Then

I. $n = \sum_{i=1}^{r} n_i$ is the number of observations obtained in all treatments.

II. $G = y = \sum_{i=1}^{r}\sum_{j=1}^{n_i} y_{ij}$ is the grand total (sum of all the observed responses of the r treatments) and $\bar{y} = \frac{G}{n}$ is the mean of all the observations.

III. $\sum_{j=1}^{n_i} y_{ij} = y_i$ is the sum of the observations in the i^{th} treatment and $\bar{y}_i = \frac{y_i}{n_i}$ is the i^{th} treatment sample mean.

9.3.1 Assumptions Required for the Analysis of Single Factor ANOVA

The assumptions underlying the ANOVA test align with the general assumptions for any parametric test:

- **Independence of observations:** The data should be collected using statistically sound sampling methods, ensuring that there are no concealed relationships among observations. If this assumption is not met, and there is a confounding variable that requires statistical control, you can employ ANOVA with blocking variables.
- **Normally distributed response variable $y_{ij} \sim N(\mu_i\ \sigma^2)$:** The values of the dependent variable have a normal distribution with mean μ_i and variance σ^2.
- **Homogeneity of variance:** The variability within each group under comparison should be roughly equal across all groups. If there are significant differences in variances among the groups, ANOVA may not be the most appropriate statistical approach for the data.

Replication increases the precision of the i^{th} treatment sample mean because,

$$\bar{y}_i = \frac{\sigma^2}{n_i}$$

The variance tends to zero as the number of replications increases.

The observation/response can be written as:

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

where $i = 1, 2, \ldots r$, $j = 1, 2, \ldots, n_i$

Where ε_{ij} is the error term distributed as $N(0, \sigma^2)$

There are k parameters in the model, *i.e.*, $\mu_1, \mu_2, \ldots, \mu_k$

Let $\mu = \frac{1}{k}\Sigma\mu_i$ be the average of all the treatment means.

The i^{th} treatment mean can be written as,

$\mu_i = \mu + (\mu_i - \mu) = \mu + t_i$ where $t_i = (\mu_i - \mu)$ is the i^{th} treatment effect.

Taking summation on both sides

$$\sum \mu_i = r\mu + \sum t_i$$

An equivalent formulation of the ANOVA modelis

$$y_{ij} = \mu + t_i + \mu_{ij} \qquad \forall\, i = 1, 2, \ldots r, j = 1, 2, \ldots, n_i$$

The above Model is referred to as the factor effects model because it is expressed in terms of the Factor effects t_i, $i = 1, 2, \ldots r$.

9.3.2 Sums of Squares

- The quantity $SST = \sum\sum(y_{ij} - \bar{y})^2$ is called the sum of squares total. For easier computation, the sum of squares total can be written as:

$$SST = \sum\sum y_{ij}^2 - \frac{G^2}{n} \quad \text{where} \quad G = \Sigma\Sigma y_{ij}$$

- The quantity $SS(tr) = \Sigma\Sigma(y_i - \bar{y})^2$ is called the sum of squares between treatments. For easier computation $SS(tr)$ can be written as:

$$SS(tr) = \sum\left(\frac{y_i^2}{n_i}\right) - \frac{G^2}{n}$$

- The sum of square total is given by $SST = \Sigma\Sigma(y_{ij} - \bar{y}_i)^2$

$$SST = SS(tr) + SSE$$

Therefore, the sum of squares within treatment is obtained as: $SSE = SST - SS(tr)$

9.3.3 Testing of Hypotheses (Overall F-test for the Equality of the Factor Level Means)

Procedure:

1. Set up null and alternative hypothesis:
   H_0 : There is no significant difference.
   H_1 : There is significant difference.
2. Compute sum of squares total (SST)

$$SST = \sum\sum y_{ij}^2 - \frac{G^2}{n} \text{ where } G = \Sigma\Sigma y_{ij}$$

3. Compute sum of squares between treatments $SS(tr)$

$$SS(tr) = \sum\left(\frac{y_i^2}{n_i}\right) - \frac{G^2}{n}$$

4. Compute sum of squares within treatment *SSE*

$$SSE = SST - SS(tr)$$

5. Compute mean sum of square between treatments *MSS*(*tr*)

$$MSS(tr) = \frac{SS(tr)}{df} = \frac{SS(tr)}{r-1}$$

6. Compute mean sum of square within treatment (*MSE*)

$$MSE = \frac{SSE}{df} = \frac{SSE}{n-r}$$

7. Compute *F*-ratio

$$F = \frac{\text{larger variance}}{\text{smaller variance}} = \frac{MSS(tr)}{MSE}$$

8. Obtain table value at corresponding to the level of significance and for degree of freedom of $(r-1, n-r)$
9. Decide whether to accept or reject the null hypothesis.

ANOVA TABLE

Source of Variation	**Sum of Squares**	**Degree of freedom**	**Means squares**	***F* Value**	$F_{a(r-1,\ n-r)}$
Between treatment	$SS(tr)$	$r-1$	$MSS(tr) = \frac{SS(tr)}{r-1}$	$F = \frac{MSS(tr)}{MSE}$	From Table
With in treatments	SSE	$n-r$	$MSE = \frac{SSE}{n-r}$		
Total	$SST = SS(tr) + SSE$	$n-1$			

Single Factor ANOVA table

Example 9.1

Given below are the yield (in kg) per acre for 5 trial plots of 4 varieties of treatments.

Treatments

Plot Name	**1**	**2**	**3**	**4**
A	42	48	68	80
B	50	66	52	94
C	62	68	76	78
D	34	78	64	82
E	52	70	70	66

Carry out an analysis of variance and state whether there is any significant difference in treatments.

y_1	y_2	y_3	y_4	y_1^2	y_2^2	y_3^2	y_4^2
42	48	68	80	1764	2304	4624	6400
50	66	52	94	2500	4356	2704	8836
62	68	76	78	3844	4624	5776	6084
34	78	64	82	1156	6084	4096	6724
52	70	70	66	2704	4900	4900	4356
$\Sigma y_1 = 240$	$\Sigma y_2 = 330$	$\Sigma y_3 = 330$	$\Sigma y_4 = 400$	$\Sigma y_1^2 = 11968$	$\Sigma y_2^2 = 22268$	$\Sigma y_3^2 = 22100$	$\Sigma y_4^2 = 32400$

1. H_0: There is no significant difference in treatments.

   H_1: There is significant difference in treatments.

2. Compute Total sum of squares (SST)

$$n_1 = 5, n_2 = 5, n_3 = 5, n_4 = 5$$

$$n = n_1 + n_2 + n_3 + n_4 = 20$$

$$SST = \Sigma\Sigma y_{ij}^2 - \frac{G^2}{n} \quad \text{Where} \quad G = \sum\sum y_{ij}$$

$$i = 1, 2, \ldots 4, \qquad j = 1, 2, \ldots, 5$$

$$SST = (11968 + 22268 + 22100 + 32400) - \frac{(240+330+330+400)^2}{20}$$

$$= 88736 - \frac{(1300)^2}{20} = 4236$$

$$SST = \Sigma\Sigma y_{ij}^2 - \frac{G^2}{n} \text{ where } G = \Sigma\,\Sigma y_{ij}$$

3. Compute sum of squares between treatments $SS(tr)$

$$SS(tr) = \sum\left(\frac{y_i^2}{n_i}\right) - \frac{G^2}{n}$$

$$SS(tr) = \frac{(240)^2}{5} + \frac{(330)^2}{5} + \frac{(330)^2}{5} + \frac{(400)^2}{5} - \frac{(240+330+330+400)^2}{20} = 2580$$

4. Compute sum of squares within treatment SSE.

$$SSE = 4236 - 2580 = 1656$$

5. Compute mean sum of square between treatments $MSS(tr)$

$$MSS(tr) = \frac{SS(tr)}{df} = \frac{2580}{3} = 860$$

6. Compute mean sum of square within treatment (MSE)

$$MSE = \frac{SSE}{df} = \frac{SSE}{n-r} = \frac{1656}{20-4} = \frac{1656}{16} = 1035$$

7. Computer F-ratio

$$F = \frac{\text{larger variance}}{\text{small variance}} = \frac{MSS(tr)}{MSE} = \frac{860}{103.5} = 8.31$$

ANOVA Table

Source of squares	**Sum of Squares of freedom**	**Degree Variation**	**Mean**	**F Value**	**$F_\alpha(r-1,n-r)$**
Between treatments	2580	3	$MSS(tr)=860$	F=8.31	If α =5% $F_{0.05}(3,16) = 3.24$
With in Between	1656	16	$MSE = 103.5$		
Total	4236	19			

As the calculated value is less than the table value, we accept the null hypothesis.

∴ There is no significant difference between the mean yield of these varieties.

9.4 TWO-FACTOR ANOVA MODEL

Imagine a scenario where a researcher intends to assess the impact of two distinct vitamins and two distinct methods of delivering these vitamins on the weight gain of chicks. In this context, we're dealing with a two-way model. We can represent the effects of the two vitamins as β_1 and β_2, while the effects of the two administration methods can be denoted as β_1 and β_2. If the researcher considers these effects to be additive, meaning there is no interaction between them, the model can be expressed as follows:

$$y_{11} = \mu + \alpha_1 + \beta_1 + \varepsilon_{11}, \quad y_{12} = \mu + \alpha_1 + \beta_2 + \varepsilon_{12}$$
$$y_{21} = \mu + \alpha_2 + \beta_1 + \varepsilon_{21}, \quad y_{22} = \mu + \alpha_2 + \beta_2 + \varepsilon_{22}$$

Or as:

$$y_{ij} = \mu + \alpha_i + \beta_j + \mu_{ij} \; i = 1, 2; j = 1, 2$$

where y_{ij} is the weight gain of the $(i, j)^{th}$ chick and ε_{ij} is a random error.

In matrix form, the model can be expressed as:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \end{bmatrix}$$

$$y = X\beta + \varepsilon$$

Procedure:

(1) Set up null and alternative hypothesis:

H_0 : There is no significant difference between columns.
: There is no significant difference between rows.
H_1 : There is significant difference between columns.
: There is significant difference between rows.

(2) Compute sum of squares total (SST)

$$SST = \Sigma\Sigma y_{ij}^2 - \frac{G^2}{n} \text{ where } G = \Sigma\Sigma y_{ij}$$

(3) Compute sum of squares of Column (SSC)

$$SSC = \frac{(\Sigma y_1)^2}{n_1} + \frac{(\Sigma y_2)^2}{n_2} + \ldots - \frac{G^2}{n}$$

Where Σy_1, Σy_2, ... represent column total

(4) Compute sum of squares of rows (SSR)

$$SSR = \frac{(\Sigma y_1)^2}{n_1} + \frac{(\Sigma y_2)^2}{n_2} + \ldots - \frac{G^2}{n}$$

Where Σy_1, Σy_2, ... represent row total

(5) Compute sum of squares errors SSE

$$SSE = SST - (SSC + SSR)$$

(6) Compute mean sum of square of columns MSC

$$MSC = \frac{SSE}{df} = \frac{SSC}{C-1}$$

(7) Compute mean sum of square of rows MSR

$$\text{MSR} = \frac{SSR}{df} = \frac{SSR}{r-1}$$

(8) Compute MSE

$$MSE = \frac{SSE}{df} = \frac{SSE}{(r-1)(c-1)}$$

(9) Compute F-ratio in respect of column

$$F_c = \frac{MSC}{MSE} = \frac{SSR}{r-1}$$

(10) Compute F-ratio in respect of column

$$F_r = \frac{MSR}{MSE}$$

(11) Obtain table value at corresponding to the level of significance.

(12) Decide whether to accept or reject the null hypothesis.

Example 9.2 Three experimenters determine the moisture content of samples of body lotion. For this purpose, each experimenter has taken 4 consignments. The results are given below:

Experimenter	**Consignments**			
	I	II	III	IV
A	9	10	9	10
B	12	11	9	11
C	11	12	10	12

Test whether there is any significant difference among consignments and among experimenters at 5% significance of level.

Solution:

	Consignments				
	I	II	III	IV	Sum of columns
A	9	10	9	10	38
B	12	11	9	11	43
C	11	12	10	12	45
Sum of rows	32	33	28	33	126

H_0 : There is no significant difference between consignments.

: There is no significant difference between experimenters.

H_1 : There is significant difference between consignments.

: There is significant difference between experimenters.

Compute sum of squares total (SST)

(1) H_0 : There is no significant difference between consignments.

There is no significant difference between experiments.

H_1: There is significant difference between consignments.

There is significant difference between experiments.

(2) Compute sum of squares total (SST)

$$SST = \sum\sum y_{ij}^2 - \frac{G^2}{n}$$

$$= 9^2 + 10^2 + 9^2 + 10^2 + 12^2 + 11^2 + 9^2 + 11^2 + 11^2 + 12^2 + 10^2 + 12^2 - \frac{126^2}{12} = 15.0$$

(3) Compute sum of squares of column (SSC)

$$SSC = \frac{(\Sigma y_1)^2}{n1} + \frac{(\Sigma y_2)^2}{n2} + \ldots - \frac{G^2}{n} = \frac{32^2}{4} + \frac{33^2}{4} + \frac{28^2}{4} + \frac{33^2}{4} - \frac{126^2}{12} = 5.67$$

(4) Compute sum of squares of rows (SSR)

$$SSR = \frac{(\Sigma y_1)^2}{n1} + \frac{(\Sigma y_2)^2}{n2} + \ldots - \frac{G^2}{n} = \frac{38^2}{4} + \frac{43^2}{4} + \frac{45^2}{4} - \frac{126^2}{12} = 6.5$$

(5) Compute sum of squares error SSE

$$SSE = SST - (SSC + SSR) = 15 - 5.67 - 6.5 = 2.83$$

(6) Compute mean sum of square of columns MSC

$$MSC = \frac{SSC}{df} = \frac{SSC}{C-1} = \frac{5.67}{3} = 1.89$$

(7) Compute mean sum of square of rows MSC

$$MSR = \frac{SSR}{df} = \frac{6.5}{2} = 3.25$$

(8) Compute MSE

$$MSE = \frac{SSE}{df} = \frac{SSE}{(r-1)(c-1)} = \frac{2.83}{3 \times 2} = 0.47$$

(9) Compute F-ratio in respect of column

$$F_c = \frac{1.89}{0.47} = 4.0$$

(10) Compute F-ratio in respect of column

$$F_r = \frac{MSR}{MSE} = \frac{3.25}{0.47} = 6.89$$

(11) If $\alpha = 5\%$, $F_{0.5}(2,15) = 5.14$

(12) If $\alpha = 5\%$, $F_{0.5}(3,15) = 4.76$

ANOVA Table

Source of Variation	**Sum of Squares**	**Degree of freedom**	**Means squares**	**F Value**	$F_\alpha(r-1, n-r)$
Between Experimenters	6.5	2	3.25	6.89	$F_\alpha(r-1, n-r)$ If $\alpha = 5\%$ $F_{0.05}(2,15)$ $= 5.14$
With in consignments	5.67	3	1.89	4.0	$F_{0.05}(3,15)$ $= 4.76$
Error	2.83	6	0.47		
Total	15	11			

The F ratio 6.89 (>5.14) does lie in the critical region at 5% level of significance, therefore, we conclude that the mean moisture content as determined by 3 experimenters are not equal. The F-ratio 4.0 <4.76 does lie in the critical region at 5% level of significance, therefore, we conclude that the moisture content of 4 consignment may not differ from each other.

9.5 FACTORIAL EXPERIMENT

Factorial experiments involve simultaneously more than one factor and each factor is at two or more levels. Several factors affect simultaneously the characteristic under study in factorial experiments and the experimenter is interested in the main effects and the interaction effects among different factors. First, we consider an example to understand the utility of factorial experiments.

Example 9.3 Suppose the yield from different plots in an agricultural experiment depends upon

1. (*i*) Variety of crop and
   (*ii*) Type of fertilizer.

Both the factors are in the control of the experimenter.

2. (*iii*) Soil fertility. This factor is not in the control of the experimenter.

In order to compare different crop varieties

- Assign it to different plots keeping other factors like irrigation, fertilizer, etc. fixed and the same for all the plots.
- The conclusions for this will be valid only for the crops grown under similar conditions with respect to the factors like fertilizer, irrigation etc.

In order to compare different fertilizers (or different dosage of fertilizers)

- Sow single crop on all the plots and vary the quantity of fertilizer from plot to plot.
- The conclusions will become invalid if different varieties of the crop are sown.

- It is quite possible that one variety may respond differently than another to a particular type of fertilizer.

Suppose we wish to compare

- Two crop varieties - a and b, keeping the fertilizer fixed and
- Three varieties of fertilizers – *A*, *B* and *C*.

This can be accomplished with two randomized block designs (RBD) by assigning the treatments at random to three plots in any block and two crop varieties at random.

The possible arrangement of the treatments may appear as follows.

bBbAbc	*aAaBaC*
bCbBbA	*aCaAaB*
bAbCbB	*aBaCaA*

With these two RBDs,

- The difference among two fertilizers can be estimated
- But the difference among the crop varieties cannot be estimated. The difference among the crop varieties is entangled with the difference in blocks.
- On the other hand, if we use three sets of three blocks each and each block having two plots, then
- Randomize the varieties inside each block and
- Assign treatments at random to three sets.
- The possible arrangement of treatment combinations in blocks can be as follows:

aBaB	*aCbC*	*bBaB*
aBbB	*bCaC*	*aBbB*
bBaB	*aCbC*	*bBaB*

Here the difference between crop varieties is estimable but the difference between fertilizer treatment is not estimable.

Factorial experiments overcome this difficulty and combine each crop with each fertilizer treatment.

There are six treatment combinations as

aA, *aB*, *aC*, *bA*, *bB*, *bC*.

Keeping the total number of observations to be 18 (as earlier), we can use RBD with three blocks with six plots each, e.g.

bAaCaBbBaAbC
aAaCbCaBbBbA
bBaBbAaCaAbC

Now we can estimate

- The difference between crop varieties and
- The difference between fertilizer treatments. Factorial experiments involve simultaneously more than one factor each at two or more levels.

If the number of levels for each factor is the same, we call it is a **symmetrical factorial experiment**. If the number of levels of each factor is not the same, then we call it as a **symmetrical or mixed factorial experiment**.

We consider only symmetrical factorial experiments.

Through the factorial experiments, we can study

- The individual effect of each factor and
- Interaction effect.

Now we consider a 2^2 factorial experiment with an example and try to develop and understand the theory and notations through this example.

General notation for representing the factors is to use capital letters, e.g., A, B, C etc. and levels of a factor are represented in small letters. For example, if there are two levels of A, they are denoted as a_0 and a_1. Similarly, the two levels of B are represented as b_0 and b_1. Another alternative representation to indicate the two levels of A is 0 (for a_0) and 1 (for a_1). The factors of B are then 0 (for b_0) and 1 (for b_1).

Note: An important point to remember is that the factorial experiments are conducted in the design of an experiment. For example, the factorial experiment is conducted as an RBD.

9.5.1 Factorial Experiments with Factors at Two Levels (2^2 Factorial Experiment)

Suppose in an experiment, the values of current and voltage in an experiment affect the rotation per minutes (rpm) of fan speed. Suppose there are two levels of current.

- 5 Ampere, call it as level 1 (C_0) and denote it as a_0
- 10 Ampere, call it as level 2 (C_1) and denote it as a_1.

Similarly, the two levels of voltage are.

- 200 volts, call it as level 1(V_0) and denote it as b_0
- 220 volts, call it as level 2(V_1) and denote it as b_1.

The two factors are denoted as A, say for current and B, say for voltage.

In order to make an experiment, there are 4 different combinations of values of current and voltage.

1. Current = 5 Ampere and Voltage = 200 Volts, denoted as $C_0 V_0 = a_0 b_0$
2. Current = 5 Ampere and Voltage = 220 Volts, denoted as $C_0 V_1 = a_0 b_1$
3. Current = 10 Ampere and Voltage = 200 Volts, denoted as $C_1 V_0 = a_1 b_0$
4. Current = 10 Ampere and Voltage = 220 Volts, denoted as $C_1 V_1 = a_1 b_1$

The responses from those treatment combinations are represented by $a_0\, b_0 = (1)$, $a_0\, b_1 = (b)$, $a_1\, b_0 = (a)$, $a_1\, b_1 = (ab)$, respectively.

Now consider the following:

I. $\frac{(C_0V_0 + C_0V_1)}{2}$: Average effect of voltage for the current level C_0:

$$\frac{a_0b_0 + a_0b_1}{2} = \frac{(1)+(b)}{2}$$

II. $\frac{(C_1V_0 + C_1V_1)}{2}$: Average effect of voltage for the current level C_1:

$$\frac{a_1b_0 + a_1b_1}{2} = \frac{(a)+(ab)}{2}$$

III. $(C_0V_1 - C_0V_0) = (b) - (1)$: indicate the simple effect (se) of voltage on rpm at current level C_0.

IV. $(C_1\, V_1 - C_1\, V_0) = (ab) - (a)$: indicate the simple effect (se) of voltage on rpm at current level C_1.

V. Main effect of B = mean (simple effect (se) of voltage on rpm at current level C_0, Simple effect (se) of voltage on rpm at current level C_1).

Main effect (me) of $B = \frac{(C_0V_1 - C_0V_0) + (C_1V_1 - C_1V_0)}{2}$

VI. Similarly, the main effect (me) of A = mean (simple effect (se) of current on rpm at voltage level V_0, simple effect (se) of current on rpm at voltage level V_1

$$= \frac{(C_1V_0 - C_0V_0) + (C_1V_1 - C_0V_1)}{2}$$

VII. The average interaction effect of voltage and current can be obtained as:

Simple effect (se) of voltage on rpm at current level C_1- simple effect (se) of voltage on rpm at current level C_0

= Average effect of voltage at different levels of current

$$= \frac{(C_1V_1 - C_1V_0) - (C_0V_1 - C_0V_0)}{2}$$

$$= \frac{(ab)-(a)}{2} - \frac{(b)-(1)}{2}$$

= Average interaction effect.

Similarly

VIII. $\frac{(C_0V_0 - C_1V_0)}{2}$: Average effect of current for the voltage level V_0:

$$= \frac{a_0b_0 + a_1b_0}{2} = \frac{(1)+(a)}{2}$$

IX. $\frac{(C_0V_1 + C_1V_1)}{2}$: Average effect of current for the voltage level V_1:

$$= \frac{a_0b_1 + a_1b_1}{2} = \frac{(b)+(ab)}{2}$$

X. Average effect of current on rpm at voltage level V_0 – Average effect of current on rpm at voltage level V_1

= Average effect of current on rpm at different levels of voltage

$$= \frac{(C_1V_1 - C_0V_1)}{2} - \frac{(C_1V_0 - C_0V_0)}{2}$$

$$= \frac{(ab)-b}{2} - \frac{a-1}{2}$$

$$= \frac{(ab)-(a)}{2} - \frac{(b)-(1)}{2}$$

= Average interaction effect

= Same as average effects of voltage at different levels of current.

= (It is expected the interaction effect of current and voltage is same as the interaction effect of voltage and current).

XI. The quantity $\frac{(C_0V_0 + C_1V_0 + C_0V_1 + C_1V_1)}{4} = \frac{(1)+(a)+(b)+(ab)}{4}$ gives the general mean effect of all the treatment combination

Example 9.4: An example of an experiment involving two factors is the application of two nitrogen levels, N_0 and N_1, and two phosphorous levels, P_0 and P_1 to a crop, with yield (lb/a) as the measured variable. The results are shown here:

Factor			**A = N level**		
	Level	$a_1 = N_0$	$a_2 = N_1$	Mean	$a_2 - a_1$
$B =$	$b_1 = P_0$	40.9	47.8	44.4	6.9 (se A, b_1)
P level	$b_2 = P_1$	42.4	50.2	46.3	7.8 (se A, b_2)
	Mean	41.6	49.0	45.3	7.4 (me A)
	$b_2 - b_1$	1.5 (se B, a_1)	2.4 (se B, a_2)	1.9 (me B)	

- The differences $a_2 - a_1$ at b_1 and $a_2 - a_1$ at b_2 are the simple effects, denoted (se A) and (se B). For example, 6.9 lb/a is the simple effect of N on yield at Po.
- The averages of the simple effects are the main effects, denoted (me A) and (me B) or simply (A) and (B). For example, 7.4 lb/a is the main effect (me) of N on yield.

  There are two ways of defining an interaction between two factors A and B:
  - If the average change in response between the levels of factor A is not the same at all levels of factor B, then an interaction exists between factors A and B.
  - The lack of additivity of factors A and B, or the non-parallelism of the mean profiles of A and B, is called the interaction of A and B.
- When we assume there is no interaction between A and B, we say the effects are additive.

Disadvantages of Factorial Design

Large number of combinations required for studying several factors at several levels and there is also a need for large-sized experiment: 7 factors at 3 levels require 2187 combinations.

Large number of factors complicates the interpretation of high order interactions.

9.6 THE TWO-WAY FACTORIAL ANALYSIS

9.6.1 The Linear Model for two-way Factorial Experiments

We have considered the situation up to now by assuming only one observation for each treatment combination, *i.e.*, no replication. If n replicated observations for each of the treatment combinations are obtained, then the expressions for the main and interaction effects can be expressed as:

$$\text{Main effect (me)of } A = \frac{(C_1V_0 - C_0V_0) + (C_1V_1 - C_0V_1)}{2n}$$

Main effect (me) of $B = \dfrac{(C_0V_1 - C_0V_0) + (C_1V_1 - C_1V_0)}{2n}$

Average Intereaction Effect (aie) $= \dfrac{(ab)^{2n} - (b)}{2n} - \dfrac{(a) - (1)}{2n}$

The linear model for a two-factor analysis is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

α_i represents the main effect of factor A_i, $i = 1, 2, \ldots, a$

β_j represents the main effect of factor B_j, $j = 1, 2, \ldots, b$

$(\alpha\beta)_{ij}$ represents the interaction of factor A level i with factor B level j

ε_{ijk} is the error associated with rep k of factor level ij, $k = 1, 2, \ldots, n$; where n observations are obtained for each treatment combinations.

When the experiments are conducted factor by factor, then much more resources are required in comparison to the factorial experiment. For example, if we conduct RBD for three-levels of voltage V_0, V_1, and V_2 and two levels of current I_0 and I_1, then to have 10 degrees of freedom for the error variance, we need

- 6 replications on voltage
- 11 replications on current.

So the total number of fans needed is 40.

For the factorial experiment with 6 combinations of 2 factors, the total number of fans required amount to 18 for the same precision.

9.6.2 ANOVA for a Two-way Factorial Design

We use the same steps for 2- Factor ANOVA that we have used for all other test statistics.

Write the alternative and null hypotheses.

3 separate set of hypotheses: one set for each F

A effect (factor 1)

B effect (factor 2)

Interaction ($A \times B$ or factor 1×2)

These are three separate ANOVA tests yielding 3 Fs that are independent, and the results are unrelated to the outcome for either of the other two.

The H_0 for a 2-factor experiment are $\alpha_i = 0$, $\beta_j = 0$, and $(\alpha\beta)_{ij} = 0$.

In ANOVA for factorial experiments the *SS* of treatments is partitioned into components for each factor and each interaction.

$$SS = SSA + SSB + SSAB + SSE$$

Where

$$SSA = (\mathrm{me}(A))^2$$
$$SSB = (\mathrm{me}(B))^2$$
$$SSAB = (aie)^2$$

And

Total sum of square (SST) is

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{G^2}{4r}$$

$$G = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$$

and

$$SSE = SST - SSA - SSB - SSAB$$

This partition is valid even when the overall F test of no differences among treatments is not significant. A significant response to A might well be lost in an overall test of significance.

The ANOVA table for the CRD factorial design is

Source	***Df***	**SS**	***MS***	***F***
Factor A	$a-1$	SSA	$MSA = SSA/(a-1)$	$F_A = MSA/MSE$
Factor B	$b-1$	SSB	$MSB = SSB/(b-1)$	$F_B = MSB/MSE$
A×B	$(a-1)(b-1)$	$SSAB$	$MSAB = SSAB/(a-1)(b-1)$	$F_{AB} = MSAB/MSE$
Error	$ab(n-1)$	SSE	$MSE = SSE/ab(n-1)$	
Total	$nab-1$	SST		

The decision rule is to reject the concerned null hypothesis when the value of the concerned F statistic

$$F_{\text{effect}} > F_{(1-\alpha)}\,(dof\,(effect),\, ab(n-1))$$

No significant interaction: multiple comparisons can be performed on the main effect means.

Significant interaction: go back to the means and analyse simple effects. Compare he means of one factor separately for each level of the other factor

xample 9.5 A farmer wants to see if there is a difference in the average height for two ew strains of hemp plants. They believe there also may be some interaction with different oil types so they plant 5 hemp plants of each strain in 4 types of soil: sandy, clay, loam and t. At $\alpha = 0.01$, analyse the data shown, using a two-way ANOVA as started below in g. 9.1. See below for raw data.

	Sandy	Clay	Loam	Silt
Strain A	60	54	80	62
	53	63	82	76
	58	62	62	55
	62	71	88	48
	57	76	71	61
Strain B	36	62	68	63
	41	61	72	65
	54	77	71	72
	65	53	82	71
	53	64	86	63

Source	SS	df	MS	F
Strain	12.1			
Soil Type	2501			
Interaction	268.1			
Error				
Total	5108			

Fig. 9.1

First, find the missing values for the ANOVA table in Fig. 9.1. The sum of squares should add up to the total in the SS column of 5108.

Subtract the 3 sum of squares that were already given in the table from the total sum of squares to find the missing error sum of squares:

$SSE = 5108 - 12.1 - 2501 - 268.1 = 2326.85108 - 12.1 - 2501 - 268.1 = 2326.8.$

Find the degrees of freedom for each category.

There are 2 types of hemp strains so $a = 2$ and the $df_A = 2 - 1 = 1$.

There are 4 soil types so $b = 4$ and the $df_B = 4 - 1 = 3$.

The interaction $df_{A \times B} = (2 - 1)(4 - 1) = 3$

There are 5 plants within each group so $n = 5$, and $df_E = 2 \times 4\,(5 - 1) = 32$.

The total df should add up to $N - 1$. There are 8 groups of 5 so there are 40 data points all together; $df_{\text{Total}} = 39$. Fill in your *SS* and df to the ANOVA table and then divide the sum of squares by df to get the mean squares. See below.

Source	SS	df	MS	F
Strain	12.1	1	12.1	
Soil Type	2501	3	833.6667	
Interaction	268.1	3	89.3667	
Error	2326.8	32	72.7125	
Total	5108	39		

Fig. 9.2

Next, find the test statistic
for by dividing the MS for each category by the MSE.
For the row effect $F_A = 12.1/72.7125 = 0.1664$;
for the column effect $F_B = 833.6667/72.7125 = 11.4652$; and
for the interaction effect $F_{(A\times B)} = 89.3667/72/7125 = 1.2290$
Add these to complete the ANOVA table.

Source	**SS**	**df**	**MS**	**F**
Strain	12.1	1	12.1	0.1664
Soil Type	2501	3	833.6667	11.4652
Interaction	268.1	3	89.3667	1.229
Error	2326.8	32	72.7125	
Total	5108	39		

Fig. 9.3

There are three hypothesis tests that we are performing: the strain of hemp plan has an effect on the mean plant height (row effect), the type of soil has an effect on the mean plant height (column effect), and the interaction effect between strain of hemp plan and soil type on the mean plant height.

Row Effect-Factor A:

H_0 : The strain of hemp plant has no effect on the mean plant height.

H_1 : The strain of hemp plant has an effect on the mean plant height.

Test Statistic: $F_A = 0.1664$.

Critical Value: Use an F-distribution with $\alpha = 0.01$ right-tail area, $df_A = 1$ and $df_E =$ 32.

The critical value is 7.4993. See Fig. 9.4.

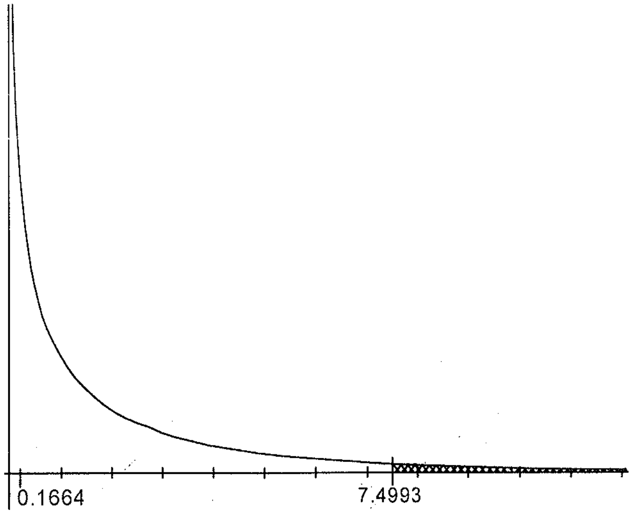


Fig. 9.4 Graph of F-distribution for the row effect (plant strain)

The test statistic of 0.1664 is less than the critical value and not in the shaded tail; therefore, do not reject H_0. There is no significant difference in the mean plant height between the two hemp strains.

Column Effect-Factor B:

H_0: The type of soil has no effect on the mean plant height.

H_1: The type of soil has an effect on the mean plant height.

Test Statistic: $F_B = 11.4652$

Critical Value: Use an F-distribution with $\alpha = 0.01$ right-tail area, $df_B = 3$ and $df_E = 32$.

The critical value is 4.4594.

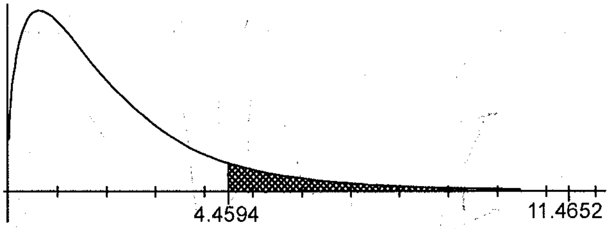


Fig. 9.5 Graph of F-distribution for the column effect (soil type)

The test statistic is greater than the critical value and is in the shaded tail; therefore, reject H0. The type of soil has a significant effect on the mean plant height.

Interaction Effect-Factor $A \times B$

H_0: There is no interaction effect between the strain of hemp and the soil type on the Mean plant height.

H_1: There is an interaction effect between the strain of hemp and the soil type on the mean plant height.

Test Statistic: $F_{A\times B} = 1.229$

Critical Value: Use an F-distribution with $\alpha = 0.01$ right-tail area, $df_{A\times B} = 3$ and $df_E = 32$.

The critical value is 4.4594.

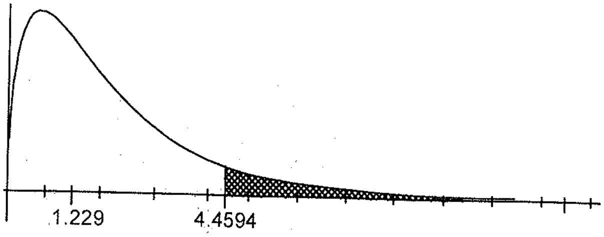


Fig. 9.6 Graph of F-distribution for the interaction effect

The test statistic is not in the shaded tail; therefore, do not reject H_0. There is no significant interaction effect between the hemp strain and the soil type on the mean plant height.

9.7 ANALYSIS OF COVARIANCE

A scientific experiment serves the purpose of uncovering unknown aspects about a set of treatments and testing specific hypotheses regarding the effects of these treatments. When the variability among experimental subjects is minimal in comparison to the differences between treatments, and the experimenter prefers not to employ a structured experimental design, many observations can be taken for each treatment, and their means can be computed. By increasing the number of observations, the variation around the mean can be reduced to the desired level.

However, in cases where there is substantial variation among observations within the same treatment, and an unlimited number of observations is not feasible, two techniques can be employed to minimize this variation:

- Utilizing an appropriate experimental design.
- Incorporating concomitant variables.

The reduction of variation using concomitant variables is achieved through a method known as analysis of covariance. If both techniques fail to effectively control experimental variability, then it becomes necessary to increase the number of replications for different treatments, or in other words, to augment the number of experimental units. This increase in the number of experimental units is done until a point is reached where adequate control over variability is achieved.

In the context of statistical modelling, the Analysis of Covariance, or ANCOVA, is a technique employed when dealing with a linear model that encompasses both quantitative and indicator (qualitative) variables.

The use of qualitative variables is particularly useful when they take on the form of indicator variables (also known as dummy variables). These indicator variables represent categories or groups within the dataset, and they can help account for differences in the response variable based on these categories.

It's important to note that quantitative variables are rich in information, as they provide a continuous range of values. As these variables could potentially influence the outcome of the experiment, they can be integrated into the model as independent variables (referred to as x's) and are then termed as covariates or concomitant variables. Then termed as covariates or concomitant variables. In contrast, indicator variables are binary (0 or 1) and can represent the presence or absence of a specific attribute or category.

As an example, consider the quantitative variable 'age.' This variable can be transformed into an indicator variable, often referred to as a dummy variable. For instance, an indicator variable for 'young' could take the value 1 if the individual is young and 0 if not. Similarly, an indicator variable for 'senior' could take the value 1 for seniors and 0 for non-seniors. These indicator variables allow us to incorporate categorical information, such as age groups, into the ANCOVA model and examine how these categories relate to the response variable.

$$D = \begin{cases} 1\ i \text{ age} \geq 17 \text{ years} \\ 0\ i \text{ page} < 17 \text{ years} \end{cases}$$

Now the following quantitative values of age can be changed into indicator variables.

Ages (years)	**Ages**
14	0
15	0
16	0
17	1
20	1
21	1
22	1

Analysis of covariance (ANCOVA) is sometimes characterized as a fusion of ANOVA and regression. The primary rationale behind utilizing covariates in an experiment is to enhance precision by reducing the error variance. In certain situations, ANCOVA can be employed to mitigate the impact of factors that the experimenter cannot effectively control, as attempting to include various levels of a quantitative variable as a full factor might render the design unwieldy. In such instances, the variable can be incorporated as a covariate, resulting in an adjustment to the dependent variable before comparing means across groups. Variables of this nature may also arise in experimental scenarios where subjects cannot be randomly assigned to treatments. In such cases, the causality implication of a designed experiment is sacrificed, and ANCOVA aligns more with descriptive model building.

For a one-way model with one covariate, the success of ANCOVA depends on the satisfaction of three key assumptions:

1. The dependent variable exhibits a linear relationship with the covariate. If this assumption holds, a portion of the error in the model is predictable and can be eliminated to reduce the error variance. This assumption can be verified by testing the null hypothesis H_0: $\beta = 0$, where β represents the slope derived from the regression of the dependent variable against the covariate. Since the estimated slope $(\hat{\beta})$ will never be exactly 0, ANCOVA will consistently yield a smaller error sum of squares compared to the equivalent ANOVA. Nonetheless, if $\hat{\beta}$ is close to 0, the slight reduction in error sum of squares may not offset the loss of a degree of freedom. This issue is more likely to arise when dealing with multiple covariates, particularly if they are strongly correlated.
2. The assumption that the groups (treatments) share the same slope is essential to assumption 1, as it implies a common slope 'β' for all 'k' groups in a one-way model with 'k' groups. To evaluate this assumption, we can test the null hypothesis H_0: $\beta_1 = \beta_2 = \beta = \beta_k$, where '$\beta_i$' represents the slope in the i^{th} group.
3. Furthermore, assumption 3 pertains to the covariate not influencing the differences among the means of the groups (treatments). If adjusting for the covariate results in a reduction of differences among the group means, the test for equality of group means would lose some of its statistical power. To verify assumption 3, one can conduct an ANOVA on the covariate to assess whether it has a significant impact on the differences between the group means.

Covariates can come in two forms, either as fixed constants (values predetermined by the researcher) or as random variables. The models here discussed deal with fixed covariates, but in practice, most of them are random. However, the estimation and testing procedures remain consistent in both scenarios, although the properties of estimators and tests exhibit some distinctions between fixed and random covariates. For instance, in cases involving fixed covariates, the test's statistical power depends on the specific values chosen for the covariates, while in situations with random covariates, the test's power relies on the population covariance matrix of the covariates.

To illustrate the application of analysis of covariance, let's consider a scenario where we want to compare the effectiveness of three different language teaching methods. We have three available classes, and each teaching method is assigned to one of these classes. Students are allowed to enrol in any of the three classes, and thus, they are not assigned randomly. There is a possibility that one class may attract a disproportionately high number of the best-performing students, making it difficult to attribute any significant difference in final grades solely to the teaching methods. Nevertheless, we can introduce previous grades or other performance metrics as covariates and subsequently compare the students' adjusted scores across the three teaching methods.

9.7.1 The Analysis of Covariance (ANCOVA) Model

In general, an analysis of covariance model can be written as:

$$y = Z\alpha + X\beta + \varepsilon$$

where Z contains 0s and 1s and α contains μ and parameters such as α_i, β_j, and $(\alpha\beta)_{ij}$ representing factors and interactions (or other effects); X contains the covariate values; and β contains coefficients of the covariates. We use $X\beta$ to represent the covariates in the model.

9.7.2 One-Way Model with One Covariate

The one-way (balanced) model can be expressed as:

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}, i = 1, 2, \ldots, k, j = 1, 2, \ldots, n$$

All kn observations can be written in the form of

$$y = Z\alpha + X\beta + \varepsilon$$

The parameter α can be estimated as:

$$\hat{\alpha} = \hat{\alpha}_0 - (Z^T Z)^{-1} Z^T X\hat{\beta}$$

$\hat{\alpha}_0 = (Z^T Z)^{-1} Z^T y$ is a solution for the normal equations for the model $y = Z\alpha + \varepsilon$ without the covariates.

Now the estimator of β is:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

where
$$S_{xx} = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2$$

$$S_{xy} = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)$$

$$S_{yy} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$$

For the analysis-of-covariance model we denote SSE as $SSE_{y.x}$

$$SSE_{y.x} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 - \frac{(\Sigma_{i=1}^{k}\Sigma_{j=1}^{n}(x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i))^2}{\Sigma_{i=1}^{k}\Sigma_{j=1}^{n}(x_{ij} - \bar{x}_i)^2}$$

which has $k(n-1)-1$ degrees of freedom. Note that the degrees of freedom of $SSE_{y.x}$ are reduced by 1 for estimation of b, since $SSE_y = S_{yy}$ has $k(n-1)$ degrees of freedom and $\frac{S_{xy}^2}{S_{xx}}$ has 1 degree of freedom. In using analysis of covariance, the researcher expects the reduction from SSE_y to $SSE_{y.x}$ to at least offset the loss of a degree of freedom.

9.7.3 Testing Hypotheses

For testing hypotheses, we assume that the $\varepsilon_{ij's}$ in are independently distributed as $N(0, \sigma^2)$. We begin with a test for equality of treatment effects.

9.7.3.1 Treatments

To test

$$H_{01} : \alpha_1 = \alpha_2 = ..., \alpha_k$$

adjusted for the covariate, we use a full-reduced model approach. The full model is in section 9.7.2 and the reduced model ($\alpha_1 = \alpha_2 = ..., = \alpha_k = \alpha$) is

$$y_{ij} = \mu + \alpha + \beta x_{ij} + \varepsilon_{ij}$$
$$y_{ij} = \mu^* + \beta x_{ij} + \varepsilon_{ij},\ i = 1, 2, ..., k, j = 1, 2, ..., n$$
$$\mu^* = \mu + \alpha$$

This is essentially the same as the simple linear regression model. SSE for this reduced model (denoted by SSE_{rd}) is given by

$$SSE_{rd} = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2 - \frac{\left(\Sigma_{i=1}^{k}\ \Sigma_{j=1}^{n}(x_{ij} - \bar{x}_i\left(y_{ij} - \bar{y}_i\right)\right)^2}{\Sigma_{i=1}^{k}\ \Sigma_{j=1}^{n}(x_{ij} - \bar{x}_i)^2}$$

which has $kn - 1 - 1 = kn - 2$ degree of freedom.

The sum of squares for testing H_{01} can be expressed as:

$$SS(\alpha|\mu, \beta) = SS(\alpha, \mu, \beta) - SS(\mu, \beta)$$

SSE_{yx} is for the full model, and SSE_{rd} is for the reduced model. They can therefore be written as:

$$SSE_{y.x} = y^T y - SS(\alpha, \mu, \beta)$$
$$SSE_{rd} = y^T y - SS(\mu, \beta)$$

Therefore;

$$SS(\alpha|\mu, \beta) = SSE_{rd} - SSE_{yx}$$

which has $kn - 2 - [k(n-1) - 1] = k - 1$ degrees of freedom. The test statistic for H_{01}: $\alpha_1 = \alpha_2 = ... = \alpha_k$ is therefore given by:

$$F = \left(\frac{SS(\alpha \mid \mu, \beta)}{k-1}\right) \Big/ \left(\frac{SSE_{yx}}{[(kn-1)-1]}\right)$$

which is distributed as $F_{[k-1,(kn-1)-1]}$, when H_{01} is true.

Hence

$$SSE_{rd} = SS(\alpha \mid \mu, \beta) + SSE_{yx}$$

ANCOVA is a valuable tool in statistical analysis, as it enables researchers to assess the impact of both quantitative and qualitative factors on a response variable while controlling for potential confounding variables. This technique is commonly used in various fields, including psychology, medicine, and social sciences, to provide a more comprehensive understanding of the relationships between variables of interest.

It helps is reducing the sum of squares due to error which in turn reflects the better model adequacy diagnostics.

PROBLEMS

1. What are the main characteristics of analysis of variance (ANOVA) models?
2. What are the main assumptions required for the analysis of single factor ANOVA?
3. Describe the procedure of Hypothesis testing for Two factor ANOVA model.
4. Three experimenters determine the moisture content of samples of body lotion. For this purpose, each experimenter has taken 4 consignments. The results are given below:

Experimenter	**Consignments**			
	I	**II**	**III**	**IV**
E_1	9	10	9	11
E_2	10	11	9	10
E_3	10	12	10	11

Test whether there is any significant difference among consignments and amongexperimenters at 5% significance of level.

5. Consider three different teaching methods (A, B, and C) and their corresponding test scores for a sample of students:

Method A: [75, 82, 78, 88, 92]
Method B: [68, 75, 80, 85, 88]
Method C: [72, 78, 84, 90, 94]

Perform a one-way ANOVA to test whether there are any significant differences in the mean test scores among the three teaching methods. Provide the ANOVA table and state your conclusion.

6. Conduct a two-way ANOVA for a study examining the effects of two factors, temperature (levels: low, medium, high) and time (levels: 1 hour, 2 hours), on the growth of plants. The data is as follows:

Low Temp: [12, 15, 18]

Medium Temp: [20, 22, 25]

High Temp: [28, 30, 35]

Perform the ANOVA and interpret the results, including any interaction effect between temperature and time.

❑❑❑

CHAPTER

10

MODEL DIAGNOSTIC

Chapter Outline

• Introduction • Checking of the Linear Relationship Between Study and Explanatory Variables • Residual Analysis • Methods for Scaling Residuals • Diagnostic Plots • Transformation and Weighting to Correct Model Inadequacies • Analytical Method for Selecting a Transformation on Study Variable (The Box-Cox Method) • Leverage and Influence Diagnostics • Variable Selection • Model Building

10.1 INTRODUCTION

The fitting of a linear regression model and the estimation of parameters, along with hypothesis testing and properties of the estimator, are contingent upon several pivotal assumptions:

- **Linearity Assumption:** The relationship between the study of a variable and the other explanatory variables is linear, at least approximately.
- **Zero Mean Error Assumption:** The error term possesses a mean of zero.
- **Homoscedasticity Assumption:** The error term maintains a constant variance.
- **Uncorrelated Errors Assumption:** The errors are uncorrelated.
- **Normality Assumption:** The errors follow a normal distribution.

The validity of these assumptions is imperative for the results to carry its meaning. When these assumptions are violated, the outcomes can be incorrect and can have significant consequences. Minor departures from these assumptions may not substantially alter the result. However, if the deviations are substantial, the obtained model could become unstable, implying that a different sample could yield an entirely different model with contradictory conclusions. Thus, it is crucial to verify these underlying assumptions before embarking on regression modelling. This information cannot be obtained solely from summary statistics such as T-statistics, F-statistics, or the coefficient of determination.

It is worth noting that these assumptions pertain to the population, while we are working with a sample. Therefore, the primary challenge lies in making inferences about the population based on a sample of data.

To assess the violation of regression assumptions, several diagnostic methods rely on the examination of model residuals using various types of graphical tools.

10.2 CHECKING OF THE LINEAR RELATIONSHIP BETWEEN STUDY AND EXPLANATORY VARIABLES

1. Case of one Explanatory Variable

When the model contains only one explanatory variable, assessing the presence of a linear relationship between y and X is straightforward. This can be achieved by examining a scatter diagram of the available data. If the scatter diagram displays a linear trend, it signifies that the relationship between y and x is indeed linear. Conversely, if the pattern observed in the scatter diagram is not linear, it suggests that the relationship between y and x takes on a nonlinear form. For example, in the following figure, a clear linear trend is evident:

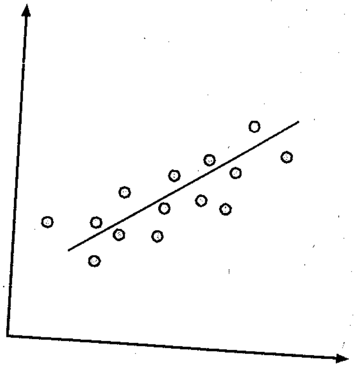


Fig. 10.1. Linear Trend

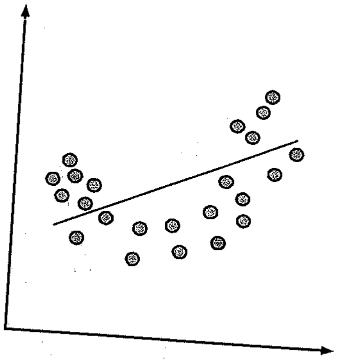


Fig. 10.2. Nonlinear Trend

2. Case of more than one Explanatory Variables

To examine the assumption of linearity between the study variable and the explanatory variables, one can utilize a scatter plot matrix of the dataset. A scatter plot matrix is an arrangement of two-dimensional plots where each cell contains a scatter diagram, except for the diagonal cells. Consequently, each combination provides insights into the relationship between a pair of variables. It offers more insights compared to just the correlation coefficient between variable pairs because it conveys information about whether the relationship is linear or nonlinear and provides a sense of how individual data points are distributed within the space. Essentially, it is a collection of scatter diagrams that depict (y versus x_1), (y versus x_2), and so on, up to (y versus x_k).

Another option to present the scatter plot is as follows:

- Display the scatter plots in the upper triangular part of the plot matrix.
- Mention the corresponding correlation coefficients in the lower triangular part of the matrix.

If there are only two explanatory variables, and the model is

$$y = x_1 \beta_1 + x_2 \beta_2 + \varepsilon,$$

then the scatterplot matrix looks like the following.

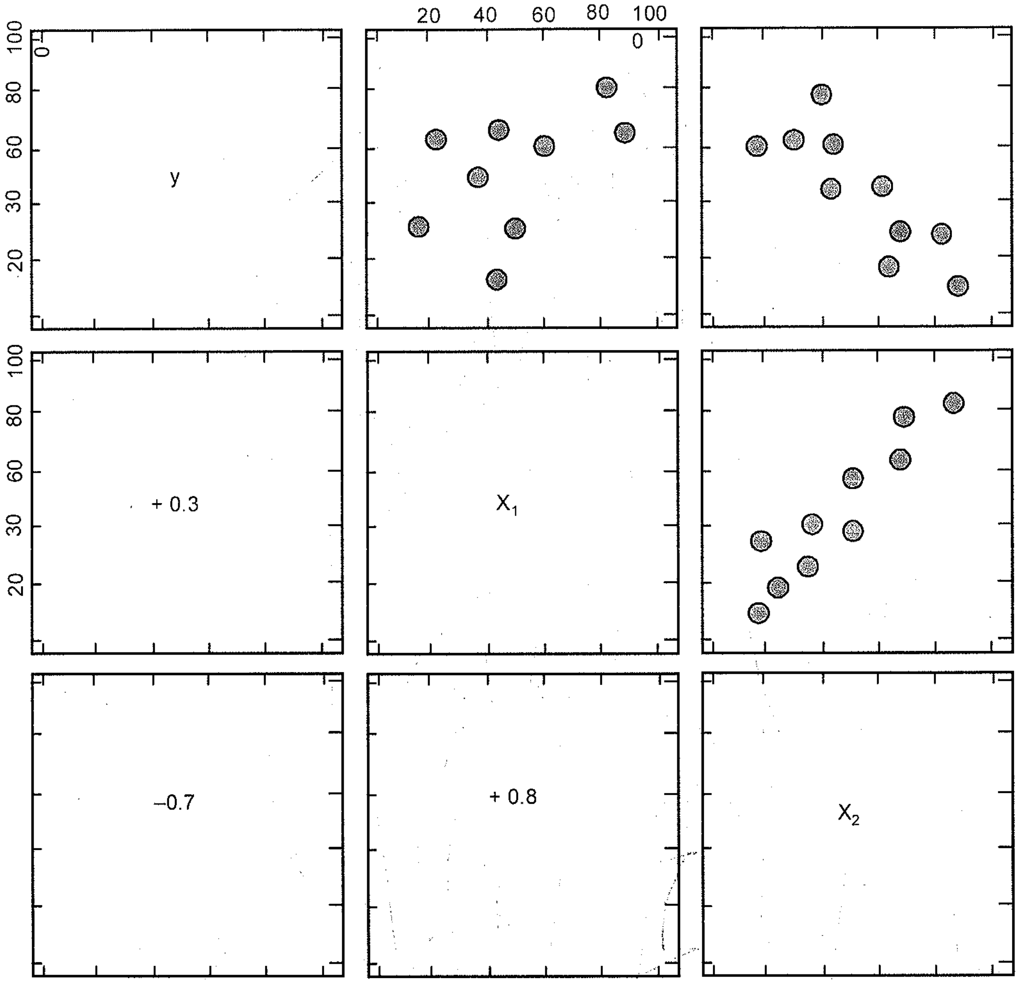


Fig. 10.3 Scatter plot

This arrangement facilitates the examination of plots and their corresponding correlation coefficients together. It's important to consider both the pair wise correlation coefficients and their corresponding scatter plots because:

- The correlation coefficient solely quantifies linear relationships.
- The correlation coefficient is non-robust, meaning that one or two outliers can significantly impact its value in the dataset.

While the presence of linear patterns in the scatter plots is reassuring, the absence of such patterns do not necessarily imply that the linear model is incorrect. Most statistical software packages offer the option to generate a scatter plot matrix, which provides a comprehensive view of all the plots. This comprehensive view often indicates whether a multiple linear regression model may offer a reasonable fit for the data.

It inote worthy that the scatter plot matrix provides information about pairs of variables, such as (y versus x_1), (y versus x_2), and so on, while the assumption of linearity pertains to the relationship between y and the set of explanatory variables jointly, represented as $(x_1, x_2, ..., x_k)$.

If some of the explanatory variables are interconnected with each other, these scatter diagrams can potentially be misleading. Therefore, alternative methods are employed to disentangle the relationships between multiple explanatory variables and a study variable in such cases.

10.3 RESIDUAL ANALYSIS

The residual is defined as the disparity between the observed value and the fitted value of the study variable. For the i^{th} observation, the residual is computed as follows:

$$\boxed{e_i = y_i - \hat{y}_i} \quad \text{where} \quad i = 1, 2, ..., n$$

Here, y_i represents an observed data point, and $\hat{y}_i$ is the corresponding predicted value obtained from the regression model.

Residuals can be seen as a measure of the deviation between the actual data and the model's predictions. They represent the amount of variability in the response variable that cannot be accounted for by the regression model.

Residuals can also be interpreted as the observed values of the model's errors. Therefore, any departure from the assumptions regarding random errors should be reflected in the residuals. Analysing the residuals aids in identifying inadequacies in the model and assessing whether the model's assumptions hold.

If the OLSE estimates the regression coefficients in the model $y = X\beta + \varepsilon$, we find that:

ordinary least square estimater (BLUE)

- Residual have zero mean as

$$E[e_i] = E[y_i - \hat{y}_i]$$

$$= E[x_i\beta + \varepsilon - x_i\hat{\beta}] = 0$$

- An approximate average variance of residuals is estimated by

$$\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-k} = \frac{\sum_{i=1}^{n} e_i^2}{n-k} = \frac{SS_{res}}{n-k} = MSS_{res}$$

- The residuals are not independent because they possess only $n - k$ degrees of freedom. This lack of independent among residuals has minimal impact on their utility for assessing model adequacy. if n is not significantly small relative to k.

10.4 METHODS FOR SCALING RESIDUALS

Scaling of residuals refers to the process of transforming or adjusting the residuals, which are the differences between the observed values and the predicted values in a statistical or regression analysis. The purpose of scaling residuals is to make them more interpretable, comparable, or suitable for specific analysis purposes. There are several common methods for scaling residuals, and each has its own benefits and use cases:

1. Standardization

This involves converting residuals to have a mean of zero and a standard deviation of one. It's achieved by subtracting the mean residual from each residual and then dividing by the standard deviation of the residuals. Since $E[e_i] = 0$ and MSS_{res} estimates the approximate average variance, so logically the scaling of residual is

$$\boxed{d_i = \frac{e_i}{\sqrt{MSS_{res}}}, \quad i = 1, 2, \ldots, n.}$$

Standardization residuals are useful for comparing the relative size of residuals across different datasets.

2. Studentized Residuals

The standardized residuals use the approximate variance of e_i as MSS_{res}. The studentized residuals use the exact variance of e_i.

In the model $y = X\beta + \varepsilon$, the ordinary least square (OLS) of β is $\hat{\beta} = (X^T X)^{-1} X^T y$, and the residual vector is

$$\boxed{e = y - \hat{y}}$$

$$\begin{aligned} &= y - X\hat{\beta} \\ &= y - X(X^T X)^{-1} X^T y \\ &= y - Hy \end{aligned}$$

$$\boxed{\begin{aligned} e &= (1 - H)y \\ H &= X(X^T X)^{-1} X^T \end{aligned}}$$

Where

We can write, $e = (1 - H)y$ as

$$\begin{aligned} e &= (1 - H)(X\beta + \varepsilon) \\ &= X\beta - HX\beta + (1 - H)\varepsilon \\ &= X\beta - X(X^T X)^{-1} X^T X\beta + (1 - H)\varepsilon \\ &= X\beta - X\beta + (1 - H)\varepsilon = (1 - H)\varepsilon \end{aligned}$$

Thus $\boxed{e = (1 - H)y = (1 - H)\varepsilon}$, so the residuals are the same linear transformation of y and ε.

$$\begin{aligned} \text{Var}(e) &= \text{Var}[(1 - H)\varepsilon] \\ &= (1 - H)^2 \,\text{Var}(\varepsilon) \\ &= (1 - H)\,\text{Var}\,(\varepsilon) \end{aligned}$$

$$\boxed{\text{Var}(e) = (1 - H)\,\sigma^2}$$

The matrix $(1 - H)$ exhibits the properties of being symmetric and idempotent, however, it is typically not diagonal. As a result, the residuals display varying variances and exhibit correlations among thernselves.

If h_{ii} is the i^{th} diagonal element of matrix $(1 - H)$ and h_{ij} is the $(i, j)^{\text{th}}$ element of matrix $(1 \ - H)$, then

$$\boxed{\begin{aligned} \text{Var}(e_i) &= \sigma^2(1 - h_{ii}) \\ \text{Cov}(e_i, e_j) &= -\sigma^2 h_{ij} \end{aligned}}$$

Since $0 \le h_{ii} \le 1$, so if MSS_{res} is used to estimate $Var(e_i)$ then

$$\boxed{\begin{aligned} \widehat{Var}(e_i) &= \hat{\sigma}^2(1 - h_{ii}) \\ &= MSS_{res}\,(1 - h_{ii}) \end{aligned}}$$

$\Rightarrow$ MSS_{res} overestimates the $Var(e_i)$.

The studentized residual is of the form

$$\boxed{r_i = \frac{e_i}{\sqrt{MSS_{res}(1 - h_{ii})}}, \; i = 1, 2, ..., n}$$

Studentized residuals are useful for identifying influential data points and outliners.

3. PRESS Residuals

The prediction sum of squares (or PRESS) is a model validation method used to assess a model's predictive ability that can also be used to compare regression models. For a data set of size n, PRESS is calculated by omitting each observation individually and then the remaining $n - 1$ observation are used to calculate a regression equation which is used to predict the value of the omitted response value.

The PRESS residuals are calculated as $e_{(i)} = y_i - \hat{y}_{(i)}$, where $\hat{y}_{(i)}$ represents the predicted value of the i^{th} response based on all observations except the i^{th} one.

The rationale for using PRESS residuals is as follows:

When the y_i is highly unusual, the regression model based on all observations may be disproportionately affected by this data point. Consequently, it can result in a $\hat{y}_{(i)}$ that closely resembles y_i, leading to small residuals e_i. This situation makes it challenging to identify any outliers effectively.

However, when we remove the i^{th} observation y_i, the prediction $\hat{y}_{(i)}$ is no longer influenced by that specific data point. As a result, the resulting residual becomes more likely to highlight the presence of an outlier.

The procedure to compute PRESS residuals involves these steps:

Exclude the i^{th} observation from the dataset.

Re-fit the regression model using the remaining $(n - 1)$ observations.

Calculate the predicted value $\hat{y}_{(i)}$ for the omitted observation.

Determine the corresponding prediction error, denoted as

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

Repeat these steps for each $i = 1, 2, \ldots, n$, where n represents the total number of observations.

These prediction errors are termed PRESS residuals because they play a crucial role in calculating the prediction error sum of squares. They are also referred to as deleted residuals

As
$$e_{(i)} = \frac{e_i}{(1-h_{ii})}$$

$$\begin{aligned}\mathrm{Var}(e_{(i)}) &= \mathrm{Var}\left(\frac{e_i}{(1-h_{ii})}\right)\\ &= \frac{1}{(1-h_{ii})^2}\,\mathrm{Var}\,(e_i)\\ &= \frac{1}{(1-h_{ii})^2}\,(1-h_{ii})\,\sigma^2\end{aligned}$$

The standardized PRESS residual is

$$\begin{aligned}\frac{e_{(i)}}{\sqrt{Var\,(e_{(i)})}} &= \frac{\dfrac{e_i}{(1-h_{ii})}}{\sqrt{\dfrac{\sigma^2}{(1-h_{ii})}}}\\ &= \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}\end{aligned}$$

$$e_{(i)} \boxed{= \frac{e_i}{\sigma}}$$

which is same as the Studentized residuals.

4. *R-student*

The studentized residual r_i is commonly utilized for outliner detection, and it involves using MSS_{res} as an estimate of σ^2 when computing r_i. This approach is known as internal scaling of the residuals because MSS_{res} is an internally derived estimate of σ^2 that comes from fitting the model to the entire dataset comprising all n observations.

The externally studentized residuals or R-student t_i are obtained by using an estimate of σ^2 based on a data set with i^{th} observation removed, say $s_{(i)}{}^2$.

$$t_i = \frac{e_i}{\sqrt{s_{(i)}^2(1-h_{ii})}},\ i = 1, 2, \ldots, n.$$

Where

$$s_{(i)}{}^2 = \frac{1}{n-k-1}\left[(n-k)\,MSS_{res} - \frac{e_i^2}{1-h_{ii}}\right]$$

In other words, a studentized deleted (or externally studentized) residual is obtained by dividing the ordinary residual by a factor that considers the mean square error calculated from the estimated model with i^{th} observation excluded, denoted as $s_{(i)}{}^2$, and the leverage, denoted as h_{ii}, as indicated in the second formula. It's important to note that the key distinction between externally studentized residuals discussed here and internally studentized residuals mentioned in the previous section lies in the fact that internally studentized residuals use the mean square error from the model based on all observations, while externally studentized residuals use the mean square error from the estimated model with the i^{th} observation omitted, $s_{(i)}^2$.

Typically, externally studentized residuals are more efficient at identifying outlying Y observations compared to internally studentized residuals. When an observation exhibits an externally studentized residual greater than 3 (in absolute value), it can be classified as an outlier. It's worth noting that in the previous section, some refer to an observation as an 'outlier' if it has an internally studentized residual exceeding 3 in absolute value. To prevent any confusion, it's essential to specify whether you're referring to internally or externally studentized residuals when identifying an observation as an outlier.

The choice of scaling method depends on the specific goals of the analysis and the characteristics of the data. Each method has its own advantages and limitations, and the selection should be made with consideration of the research objectives and the properties of the residuals.

10.5 DIAGNOSTIC PLOTS

Analysing residuals graphically is a highly effective method for assessing the suitability of a regression model's fit and validating its underlying assumptions. Different types of graphics are available for examining various assumptions, and modern regression software can generate these graphics. It's advisable to create plots for both the original residuals and scaled residuals. Usually, studentized residuals are preferred for plotting since they exhibit constant variance.

10.5.1 Normal Probability Plot

The assumption of normally distributed errors is crucial for the validity of hypothesis testing, confidence intervals, and prediction intervals. While minor deviations from normality may not significantly impact the model, substantial non-normality can be problematic. Normal probability plots are useful for assessing the assumption of normal distribution. If errors originate from a distribution with thicker and heavier tails than the normal distribution, it can lead to outliers that exert undue influence on the least-squares fit. In such instances, alternative estimation techniques, such as robust regression methods, should be explored.

The normal probability plot is a plot of the ordered standardized residuals versus the so-called normal scores. The normal scores are the cumulative probability

$$P_i = \frac{i - \frac{1}{2}}{n}, \quad i = 1, 2, \ldots, n$$

If the residuals $e_1, e_2, \ldots, e_n$ are ordered and ranked in increasing order as:

$$e_{[1]} < e_{[2]} < \ldots < e_{[n]}$$

Next, the $e_{[i]}$'s are graphed against P_i in what is known as a normal probability plot. When the residuals follow a normal distribution, the ordered residuals should closely align with the ordered normal scores. Consequently, the data points on this plot should cluster around a straight line with an intercept of zero and a slope of one. These values correspond to the mean and standard deviation of standardized residuals.

The rationales behind plotting $e_{[1]}$ against $P_i = \dfrac{i - \frac{1}{2}}{n}$ is as follows:

- Divide the whole unit area under the normal curve into n equal areas.
- We have a sample of size n data sets.
- We might 'except' that one observation lies is each section, so marked out.
- The first section has one point, so the cumulative probability is $P_1 = \dfrac{1}{n}$. Second section has one point, so cumulative probability up to second section is $P_2 = \dfrac{1}{n} + \dfrac{1}{n} = \dfrac{2}{n}$ and so on.
- Then i^{th} ordered residual observation is plotted against the cumulative area to the middle of i^{th} section, which is $\dfrac{i - \frac{1}{2}}{n}$.

- The factor $\frac{1}{2}$ is used for end correction as all the observations scattered inside the stripe are assumed to be concentrated at the midpoint of the stripe.

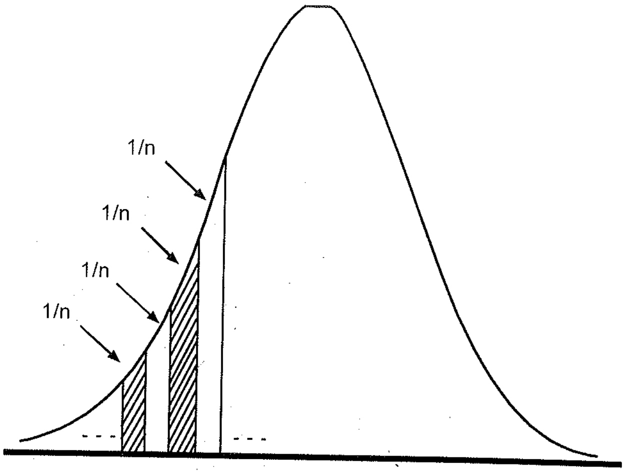


Fig. 10.4 Normal probability curve

Different softwares are used for a different criterion. For example, BMDP uses

$$P_i = \frac{i - \frac{1}{3}}{n + \frac{1}{3}}, \quad i = 1, 2, ..., n$$

which produces de-trended normal probability plots from which slope is removed.

Minitab uses

$$P_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}, \; i = 1, 2, ..., n \text{ and converts to a normal score.}$$

Such differences are not crucial in real use.

The straight line is usually determined visually with emphasis on the central values rather than the extremes. A substantial departure from a straight line indicates that the distribution is not normal.

Sometimes, the normal probability plots are constructed by plotting the ranked residuals $e_{[i]}$ against the expected normal value $\phi^{-1}\left(\frac{i - \frac{1}{2}}{n}\right)$.

where ϕ denotes the standard normal cumulative distribution. This follows from the fact that

$$E[e_{[i]}] = \phi^{-1}\left(\frac{i-\frac{1}{2}}{n}\right)$$

Graphic patterns can be interpreted as follows.

- Figure 10.5 shows an ideal normal probability plot. Points lie approximately on the straight line and indicate that the underlying distribution is normal.

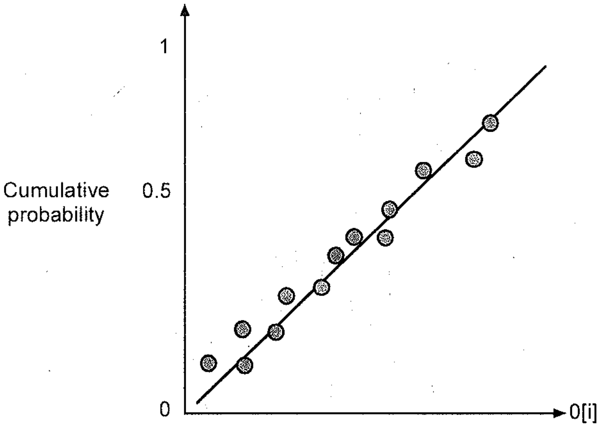


Fig. 10.5 Ideal normal probability plot

- Figure 10.6 has a step upwards and downwards sloping curves at both extremes. This indicates that the underlying distribution is heavy-tailed, i.e., the tails of the underlying distribution are thicker than the tails of a normal distribution.

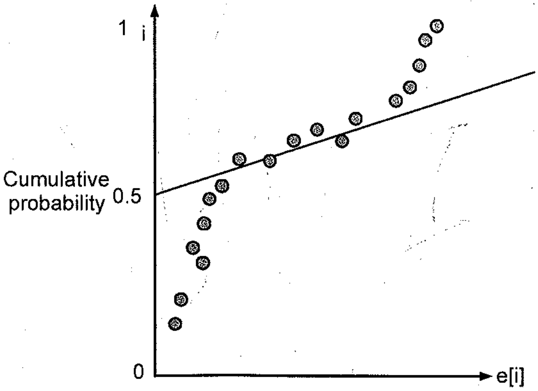


Fig. 10.6 Heavy tailed distribution

- Figure 10.7 has flattening at the extremes for the curves. This indicates that the underlying distribution is light tailed, *i.e.*, the tails of the underlying distribution are thinner than the tails of normal distribution.

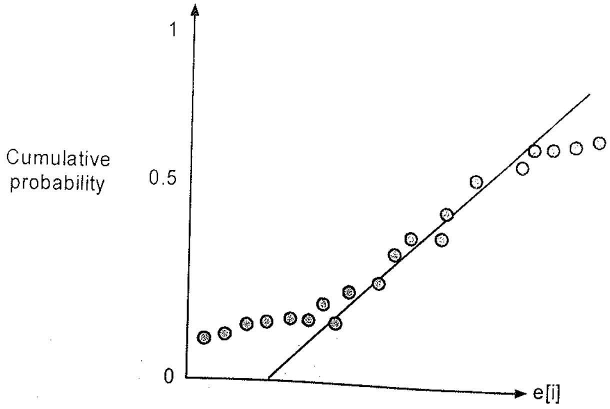


Fig. 10.7 Light tailed distribution

- Figure 10.8 has a sharp change in the direction of the trend in an upward direction from the mid. This indicates that the underlying distribution is positively skewed.

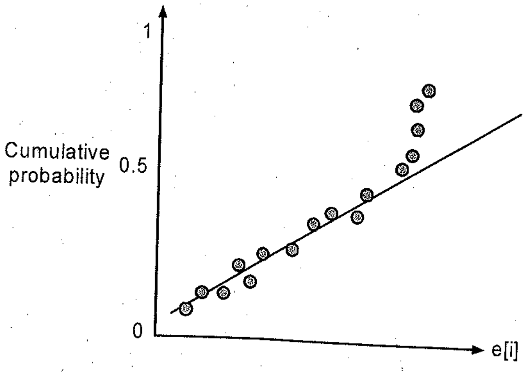


Fig. 10.8 Positively skewed distribution

- Figure 10.9 has a sharp change in the direction of the trend in the downward direction from the mid. This indicates that the underlying distribution is negatively skewed.

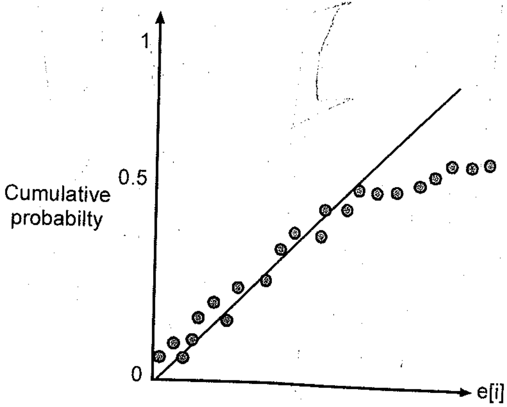


Fig. 10.9 Negatively skewed distribution

Some experience and expertise are required to interpret the normal probability plots because the samples taken from a normal distribution will not plot precisely as a straight line.

- Small sample sizes ($n \leq 16$) often produce normal probability plots that deviate substantially from linearity.
- Larger sample sizes $n \geq 32$ produce plots which are much better behaved.
- Usually, about $n = 20$ is required to produce stable and easily interpretable normal probability plots.
- If residuals are not from a random sample, normal probability plots often exhibit no unusual behaviour even if the disturbances (ε_i) are not normally distributed. Such residuals are often remnants of a parametric estimation process and are linear combinations of the model errors (ε_i).
- Thus, the fitting the parameters tends to destroy the evidence of nonnormality in the residuals. Consequently, we cannot rely on the normal probability plots to detect the departures from normality.
- Commonly seen defect found is normal probability plots is the occurrence of one or two large residuals. Sometimes, this is an indication that the corresponding observations are outliers.

10.5.2 Plots of Residuals against the Fitted Value

Plotting residuals (e_i) or any of the scaled residuals (d_i, r_i, t_i) against their corresponding predicted values $(\hat{y}_i)$ can be a useful technique for identifying various common model shortcomings. Interpretations of plots of $\hat{y}_i$ versus e_i can be summarized as follows:

(a) When the plot exhibits a pattern where the residuals are confined within a horizontal band (with fluctuations occurring more or less randomly within the band), it indicates the absence of apparent model defects. Figure 10.10 shows the residual plot for absence of model defects.

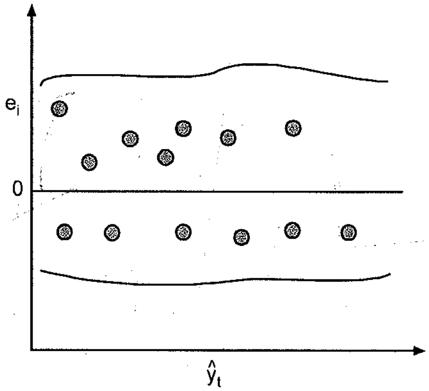


Fig. 10.10 Absence of model defects

(b) If plot is such that the residuals can be contained in an outward opening funnel then such pattern indicates that the variance of errors is not constant, but it is an increasing function of y.

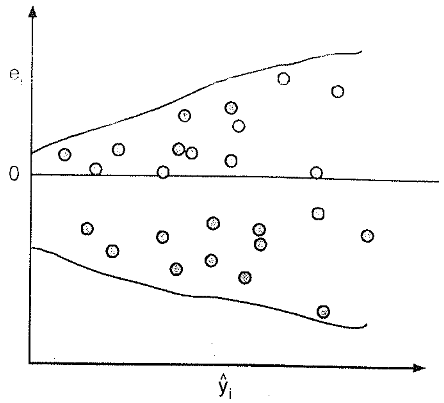


Fig. 10.11 Residual plot for varying variance as an increasing function of y

(c) If plots are such that the residuals can be accommodated in an inward opening funnel, then such a pattern indicates that the variance of errors is not constant, but it is a decreasing function of y.

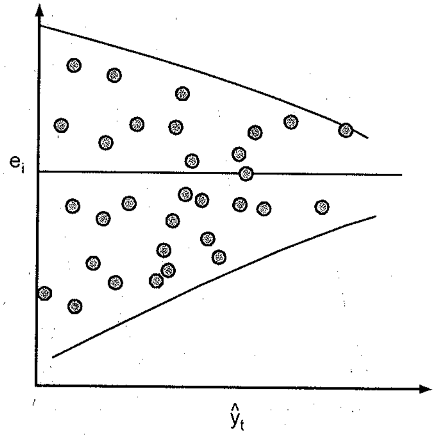


Fig. 10.12 Residual plot for varying variance as an decreasing function of y

(d) If the plot exhibits a double-bow pattern where the residuals can be fitted within it, this suggests that the variance of errors is not consistent, and y lies within the range of 0 to 1. In such a scenario, y might follow a Binomial distribution. Notably, the variance of a Binomial proportion is higher when it is close to 0.5 compared to when it is close to 0 or 1. Consequently, this implies that the assumed relationship between y and the predictor variables X is nonlinear.

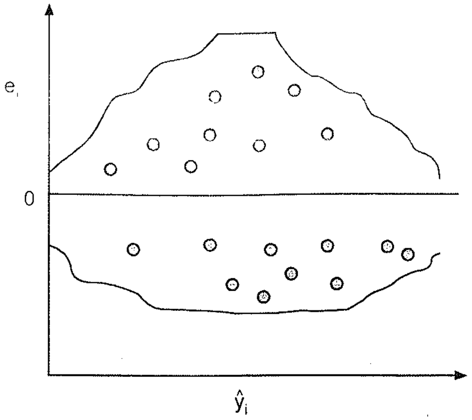


Fig. 10.13 Residual plot for non-linear relationship between y and x

The usual approach to deal with such inequality of variances is to apply a suitable transformation to either the explanatory variables or the study variable or use the method of weighted least squares. In practice, transformations on study variable are generally employed to stabilize the variance.

(e) If the plot shows that the residuals are confined within a curved pattern, it suggests the presence of nonlinearity in the model. This implies that the assumed relationship between y and the predictor variables X is nonlinear. Additionally, this could indicate the need to incorporate additional explanatory variables into the model. For instance, including a squared error term might be necessary. Transformations applied to the explanatory variables and/or the study variable could also be beneficial in addressing these situations.

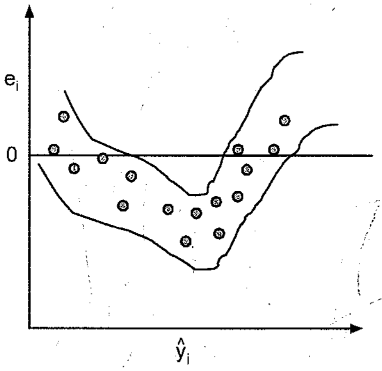


Fig. 10.14 Residual plot for non-linear relationship between y and x and need of additional explanatory variables

Important Note: When plotting residuals against the predicted values $\hat{y}_i$ it's possible to identify one or more exceptionally large residuals, which may represent potential outliers. Particularly, if these large residuals appear at the extreme values of $\hat{y}_i$ it could suggest that either the variance is not consistent or that the true relationship between y and X is nonlinear. Before classifying these points as outliers, it's crucial to investigate these possibilities thoroughly.

10.5.3 Plots of Residuals Against Explanatory Variable

Plotting of residuals against the corresponding values of each explanatory variable can also be helpful.

We proceed as follows:

- Consider the residual on Y-axis and values of the jth explanatory variable x_{ij}'s, (1, 2, ...,) on X-axis. This is the same way as we have plotted the residuals against $\hat{y}_i$, in place of $\hat{y}_i$'s now we consider x_{ij}'s.
- Interpretation of the plots is the same as in the case of plots of residuals versus $\hat{y}_i$. This is as follows. If all the residuals are contained in
  - a horizontal band, and the residuals fluctuate more or less in a random fashion within this band, then it is desirable, and there are no visible model defects.
  - an outward opening funnel shape or inward opening funnel shape, then it indicates that the variance is non constant.
  - A double bow pattern or nonlinear pattern then it indicates the assumed relationship between y and x_j is not correct.

10.6 TRANSFORMATION AND WEIGHING TO CORRECT MODEL INADEQUACIES

Graphical techniques are valuable for identifying deviations from fundamental assumptions in regression analysis. Next, we will explore methods and techniques for constructing models by transforming data when certain assumptions are violated.

10.6.1 Variance Stabilizing Transformations

In the context of regression analysis, we typically assume that the variance of errors remains constant, i.e., $Var(\varepsilon_i) = \sigma^2$. However, when this assumption is violated, it often occurs because the study variable follows a probability distribution where the variance is functionally related to the mean.

For instance, if the response variable (y) in the regression model follows a Poisson distribution, the variance of y is equal to its mean. As the mean of y is connected to the explanatory variable x, the variance of y becomes proportional to x in such cases. In these situations, employing variance-stabilizing transformations becomes beneficial.

In another scenario, if y represents a proportion, meaning it falls within the range 0 to 1 ($0 \leq y \leq 1$), then the variance of y is proportional to $E(y)[1 - E(y)]$. In such instances, applying a variance-stabilizing transformation proves to be helpful.

Various variance-stabilizing transformations, ranked in terms of their effectiveness, include:

Table 10.1 Various variance-stabilizing transformations

Relation of σ^2 to $E(y)$	**Transformation**
$\sigma^2 \propto$ constant	$y^* = y$ (No transformation)
$\sigma^2 \propto E(y)$	$y^* = \sqrt{y}$ (Poisson Data)
$\sigma^2 \propto E(y)[1 - E(y)]$	$y^* = \sin^{-1}(\sqrt{y})$ (Binomial proportion $0 \leq y \leq 1$)
$\sigma^2 \propto [E(y)]^2$	$y^* = \ln(y)$
$\sigma^2 \propto [E(y)]^3$	$y^* = \dfrac{1}{\sqrt{y}}$
$\sigma^2 \propto [E(y)]^4$	$y^* = \dfrac{1}{y}$

After making a suitable transformation, use y^* as a study variable in the respective case. The strength of a transformation depends on the amount of curvature observed in the relationship between the study variable and the explanatory variable. These transformations vary from relatively mild to relatively potent, with the square root.

Transformation falling on the milder end and the reciprocal transformation on the stronger end. Generally, a mild transformation is suitable when the range between the minimum and maximum values is not substantial (e.g., $y_{max}/y_{min} < 2$ or 3) and exerts minimal influence on the curvature. Conversely, when there is a considerable difference between the minimum and maximum values, a strong transformation is necessary, significantly impacting the analysis.

In cases of non-constant variance, the Ordinary Least Squares Estimator (OLSE) remains unbiased but loses its minimum variance property. After transforming the study variable as y^*, the predicted values exist in the transformed scale. Frequently, it becomes essential to convert these predicted values back to their original units (y). When applying an inverse transformation directly to the original values, it yields an estimate of the median of the study variable's distribution rather than the mean. Therefore, proper caution is required while doing so.

Confidence intervals and prediction intervals can be converted directly from one scale to another. This is because interval estimates represent percentiles of the distribution, and percentiles remain unaffected by the transformation. It is note worthy that the resulting intervals may or may not be the shortest possible intervals.

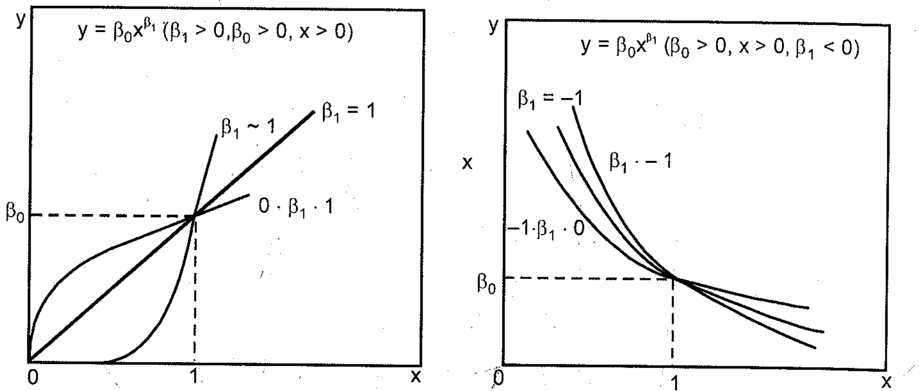
10.6.2 Transformations to Linearize the Model

The fundamental assumption in linear regression analysis asserts that the relationship between the response study variable and the predictor explanatory variables is linear. When this assumption is violated, several checks can be performed, including scatterplot matrices, scatter diagrams, partial regression plots, and lack-of-fit tests.

In certain instances, a nonlinear model can be transformed into a linear one through an appropriate transformation. These nonlinear models are referred to as intrinsically or transformable linear models. The advantage of converting a nonlinear function into a linear one lies in the fact that statistical tools have been specifically developed for linear regression models. These tools encompass exact hypothesis tests, confidence interval estimation, and more. By transforming a nonlinear function into a linear one, all of these statistical tools can be readily applied, eliminating the need for separate development.

Some linearizable functions are as follows:

1. If the curve between y and x is like as Fig. 10.15:



(a) (b)

Fig. 10.15 (a) Function $\beta_0 x^{\beta_1}$ ($\beta_1 > 0$, $\beta_0 > 0$, $x > 0$) and (b) Function $\beta_0 x^{\beta_1}$ ($\beta_0 > 0$, $x > 0$, $\beta_1 < 0$)

Then, the potentially linearizable function takes the form of:

$$y = \beta_0 x^{\beta_1}$$

To linearize this function, we can use the following transformation:

Let $y^* = ln(y)$ and $x^* = ln(x)$, which means taking the natural logarithm on both sides of the model:

$$ln(y) = ln(\beta_0) + \beta_1 ln(x)$$

This can be rewritten as:

$$ln(y) = \beta_0{}^* + \beta_1 x^*$$

Here, $\beta_0^* = ln(\beta_0)$, and the model has now become a linear one. It's important to note that the parameter β_0 changes to $\log(\beta_0)$ in the transformed model.

2. If the relationship between y and x follows the pattern as given in Figure 10.16 described as follows,

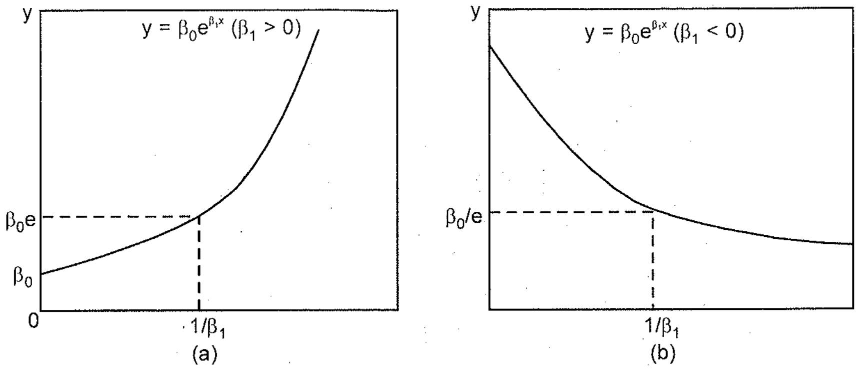


Fig. 10.16 (a) Function $\beta_0 e^{\beta_1 x}$ for ($\beta_1 > 0$) and (b) Function $\beta_0 e^{\beta_1 x}$ for ($\beta_1 < 0$)

Then the function that can be linearized is of the form:

$$y = \beta_0 \exp(\beta_1 x)$$

To linearize this function, we can apply the logarithm (*ln*) to both sides:

$$ln(y) = \log(\beta_0 \exp(\beta_1 x))$$

This simplifies to:

$$\log(y) = \beta_0{}^* + \beta_1 x$$

Here, $\beta_0^* = \log(\beta_0)$, and the model has been transformed into a linear one. It's worth noting that the intercept term β_0 changes to $\log(\beta_0)$ in the transformed model.

3. When the relationship between y and x exhibits the following pattern Figure 10.16, the function that can be linearized takes the form:

$$y = \beta_0 + \beta_1 \log(x)$$

This can be rewritten as:

$$y = \beta_0 + \beta_1 x^*$$

Here, we've applied the transformation $x^* = \log(x)$ to linearize the model.

The expression can be expressed as:

$$y = \beta_0 + \beta_1 x^*$$

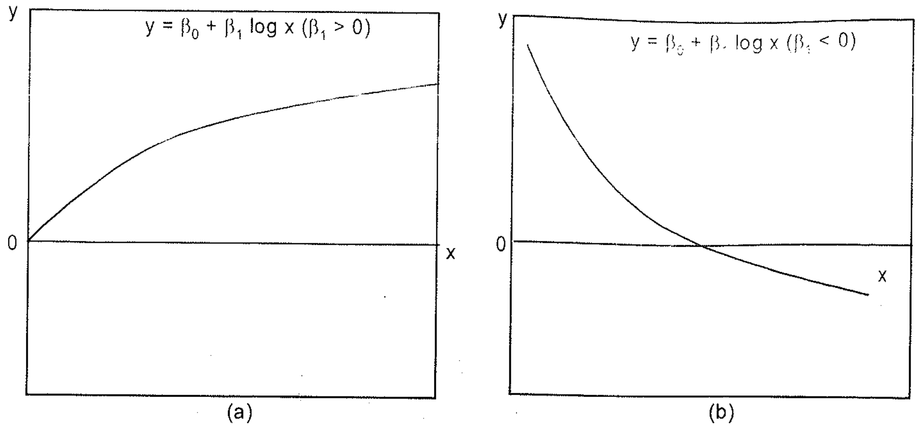


Fig. 10.17 (a) Function $\beta_0 + \beta_1 \log(x)$ for ($\beta_1 > 0$) and (b) Function $\beta_0 + \beta_1 \log(x)$ for ($\beta_1 < 0$)

4. If the relationship between y and x takes on the following form as given in Fig. 10.18:

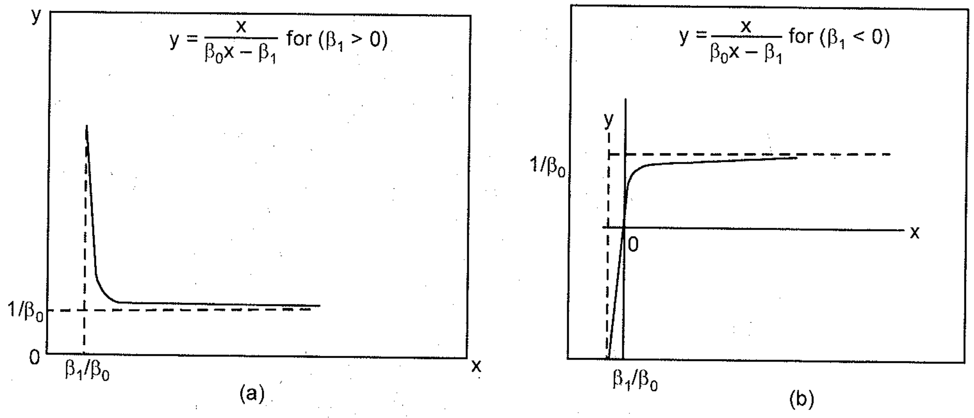


Fig. 10.18 (a) Function $\frac{x}{\beta_0 x - \beta_1}$ for ($\beta_1 > 0$) and (b) Function $\frac{x}{\beta_0 x - \beta_1}$ for ($\beta_1 < 0$)

$$y = \frac{x}{\beta_0 x - \beta_1}$$

This can be rewritten as follows:

$$\frac{1}{y} = \beta_0 - \frac{\beta_1}{x}$$

Using the transformation $y^* = 1/y$ and $x^* = -1/x$, the model becomes:

$$y^* = \beta_0 + \beta_1 x^*$$

This transformation results in a linear model. It's important to note that the form of ε (error term) may also change when such transformations are applied. For instance, in the case of:

$$y = \beta_0 \, exp(\beta_1 x) \, \varepsilon$$

When taking the natural logarithm on both sides, we get:

$$ln(y) = ln(\beta_0) + \beta_1 x + ln(\varepsilon)$$

This can be expressed as:

$$y^* = \beta_0^* + \beta_1 x + \varepsilon^*$$

Where ε^* represents the transformed error term.

- This suggests that the original model's multiplicative error follows a log-norma distribution in the transformed model. Often, we overlook this aspect and persis in assuming that the random errors continue to conform to a normal distribution In such instances, it is essential to verify the validity of these assumptions b examining the residuals from the transformed model.
- It's important to note that when these transformations are applied, the Ordinar Least Squares Estimator (OLSE) exhibits the desired properties with respect t the transformed data rather than the original data.

10.7 ANALYTICAL METHODS FOR SELECTING A TRANSFORMATION ON STUDY VARIABLE (THE BOX-COX METHOD)

Suppose that the normality and/or constant variance of the study variable y can be rectifie through a power transformation on y. This entails transforming y as y^λ, where λ is th parameter that needs to be determined. For instance, when $\lambda = 0.5$, the transformatio corresponds to taking the square root of y, and this transformed variable, $y^{0.5}$, is used i place of y in the study.

Now, in the linear regression model, we have parameters β, σ^2, and λ. The Box an Cox method provides a procedure for concurrently estimating both λ and the mode parameters using the maximum likelihood method.

It's worth noting that as λ approaches zero, y^λ tends towards 1. Consequently, there i a problem when $\lambda = 0$ because this would result in all observations of y becoming equal t unity. It is meaningless for all observations of the study variable to be constant. Therefor there is a discontinuity at $\lambda = 0$. One approach to address this issue is to use $\frac{y^\lambda - 1}{\lambda}$ as th study variable. Notably, as $\lambda \to 0$, $\frac{y^\lambda - 1}{\lambda} \to ln\, y$. So, a potential solution is to use th transformed study variable as follows:

$$\boxed{y^{(\lambda)} = W = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(y) & \text{for } \lambda = 0 \end{cases}}$$

The aim of the Box-Cox transformations is to ensure the usual assumptions for Linear Model hold $y^{(\lambda)} \sim N(X\beta, \sigma^2 I)$. Clearly not all data could be power-transformed to Normal. Draper and Cox (1969) studied this problem and conclude that even in cases that no power-transformation could bring the distribution to exactly normal, the usual estimates of λ will lead to a distribution that satisfies certain restrictions on the first 4 moments, thus will be usually symmetric.

The main objective in the analysis of Box-Cox transformation model is to make inference on the transformation parameter β, and Box and Cox (1964) considered two approaches.

- The first approach is to use the Maximum Likelihood method. This method is commonly used since it's conceptually easy and the profile likelihood function is easy to compute in this case. Also it's easy to obtain an approximate CI for λ because of the asymptotic property of MLE.

Here we assume that the transformed response $y^{(\lambda)} \sim N(X\beta, \sigma^2 I)$. We observe the design matrix X and the raw data y, and the model parameters are $(\lambda, \beta, \sigma^2)$. The density function of $y^{(\lambda)}$ can be written as:

$$\boxed{f(y^{(\lambda)}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[\frac{(y^{(\lambda)} - X\beta)'(y^{(\lambda)} - X\beta)}{2\sigma^2}\right]}$$

As the parameter λ varies, the value of $y^{(\lambda)}$ undergoes significant changes. Consequently, determining the optimal value of λ becomes a challenging task. If different analysts arrive at different λ values, it would result in fitting different models. Consequently, comparing the models with varying λ values might not be appropriate. Therefore, it is advisable to explore an alternative approach.

$$\boxed{y(\lambda, g) = \frac{y^{(\lambda)}}{g^{\lambda-1}}}$$

where g is the geometric mean of the response vector y.

$$\boxed{g = \left(\prod_{i=1}^{n} y_i\right)^{\frac{1}{n}}}$$

Let $J(y \to W)$ be the Jacobian of the transformation from y to $y^{(\lambda)}$, then the density for y (which is also the likelihood for the whole model) is

$$L(\lambda, \beta, \sigma^2 | y, X) = f(y) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{(y^{(\lambda)} - X\beta')(y^{(\lambda)} - X\beta)}{2\sigma^2}\right] J(y \to W)$$

Let $y = (y_1, y_2, ..., y_n)$, $W = (W_1, W_2, ..., W)'$.

Note that if $W_1 = \dfrac{y_1^{\lambda} - 1}{\lambda}$ then

$$\frac{\partial W_1}{\partial y_1} = \lambda \frac{y_1^{\lambda-1}}{\lambda} = y_1^{\lambda-1}$$

$$\frac{\partial W_1}{\partial y_2} = 0$$

$$\frac{\partial W_i}{\partial y_i} = \begin{cases} y_i^{\lambda-1} \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

The Jacobian of transformation is given by

$$J(y_i \to W_i) = \frac{\partial y_i}{\partial W_i} = \frac{1}{\frac{\partial W_i}{\partial y_i}} = \frac{1}{y_i^{\lambda-1}}$$

$$J(W \to y) = \begin{vmatrix} \frac{\partial W_1}{\partial y_1} & \frac{\partial W_1}{\partial y_1} & \cdots & \frac{\partial W_1}{\partial y_n} \\ \frac{\partial W_2}{\partial y_1} & \frac{\partial W_2}{\partial y_2} & \cdots & \frac{\partial W_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial W_n}{\partial y_1} & \frac{\partial W_n}{\partial y_2} & \cdots & \frac{\partial W_n}{\partial y_n} \end{vmatrix} = \begin{vmatrix} y_1^{\lambda-1} & 0 & 0 & \cdots & 0 \\ 0 & y_2^{\lambda-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & y_n^{\lambda-1} \end{vmatrix}$$

$$= \prod_{i=1}^{n} y_i^{\lambda-1} = \left(\prod_{i=1}^{n} y_i\right)^{\lambda-1}$$

$$J(y \to W) = \frac{1}{J(W \to y)} = \left(\frac{1}{\prod_{i=1}^{n} y_i}\right)^{\lambda-1}$$

Since, this is a Jacobian approach that when we want to transform the whole vector y to a whole vector W. If an individual y_i is to be transform into W_i, then take its geometric mean as:

$$J(y_i \to W_i) = \left(\frac{1}{\left(\prod_{i=1}^{n} y_i\right)^{\frac{1}{n}}}\right)^{\lambda-1}$$

The quantity $J(y \to W) = \left(\frac{1}{\prod_{i=1}^{n} y_i}\right)^{\lambda-1}$ ensures that unit volume is preserved moving from the set of y_i to the set of W_i. This is a factor which scales and ensures that the residual sum of squares of obtained from different values of λ can be compared.

To obtain the *MLE* from the last likelihood equation, we observe that for each fixed λ, the likelihood equation is proportional to the likelihood equation for estimating (β, σ^2) for observed $y^{(\lambda)}$. Here we are considering $y^{(\lambda)} = X\beta + \varepsilon$.

$$L[y^{(\lambda)}] = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[\frac{\Sigma_{i=1}^{n}\varepsilon_i^2}{2\sigma^2}\right]$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[\frac{\varepsilon'\,\varepsilon}{2\sigma^2}\right]$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[\frac{(y^{(\lambda)} - X\beta)'\,(y^{(\lambda)} - X\beta)}{2\sigma^2}\right]$$

$$lnL[y^{(\lambda)}] = -\frac{n}{2}\ln\sigma^2 - \left[\frac{(y^{(\lambda)} - X\beta)'\,(y^{(\lambda)} - X\beta)}{2\sigma^2}\right]$$

Solving $$\frac{\partial \ln L\left[y^{(\lambda)}\right]}{\partial\beta} = 0$$

$$\frac{\partial \ln L\left[y^{(\lambda)}\right]}{\partial\sigma^2} = 0$$

gives the maximum likelihood estimators.

$$\boxed{\begin{aligned} \hat{\beta}(\lambda) &= (X^T X)^{-1} X^T y^{(\lambda)} \\ \hat{\sigma}^2(\lambda) &= \frac{1}{n} y^{(\lambda)T}[1 - X(X^T X)^{-1} X^T] y^{(\lambda)} = \frac{y^{(\lambda)T}\bar{H}y^{(\lambda)}}{n} \end{aligned}}$$

For a given value of λ.

where $$\bar{H} = [1 - X(X^T X)^{-1} X^T]$$

Substituting these estimate $(\hat{\beta}(\lambda), \hat{\sigma}^2(\lambda))$ we could obtain the profile log likelihood (i.e., the likelihood function maximized over β, σ^2) for λ alone.

$$\boxed{l_p(\lambda) = l(\lambda \mid y, X, \hat{\beta}(\lambda), \hat{\sigma}^2(\lambda)) = C - \frac{n}{2}\ln(\hat{\sigma}^2(\lambda)) + (\lambda - 1)\sum_{i=1}^{n} \ln y_i}$$

Therefore, we have

$$l_p(\lambda) = C - \frac{n}{2}\ln(\hat{\sigma}^2(\lambda)) + (\lambda - 1)\sum_{i=1}^{n} \ln y_i (\lambda - 1)\sum_{i=1}^{n} \ln y_i$$

Therefore, $$l_p(\lambda) = C - \frac{n}{2}\ln(SSR(\lambda))$$

Where SSR is sum of square of residuals divided by n from fitting the linear model $y(\lambda,g) \sim N(X\beta,\sigma^2 I)$. So, to maximize the profile log-likelihood, we only need to find a λ that minimizes.

$$SSR(\lambda) = \frac{1}{n} y^T\,(y, g)[1 - X(X^T X)^{-1}X^T]\, y(\lambda, g)$$

The function $-\frac{n}{2}\ln(SSR(\lambda))$ is called as the Box-Cox objective function.

Let $\lambda_{\max}$ as the value of λ that minimizes the Box-Cox objective function. Then, under reasonably broad conditions, for any other λ the expression:

$$\frac{n}{2}\ln(SSR(\lambda)) - \frac{n}{2}\ln(SSR(\lambda_{\max}))$$

has a chi-squared distribution with one degree of freedom ($\chi^2(1)$). This outcome is derived from the behaviour exhibited by the likelihood ratio statistics in large samples.

10.8 LEVERAGE AND INFLUENCE DIAGNOSTICS

The positioning of data points within the x-space can hold significant importance when determining regression coefficients. Let's consider a scenario illustrated as follows:

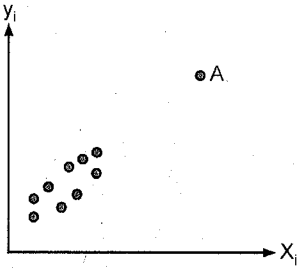


Fig. 10.19 Leverage point

In this figure, point A stands out as an outlier within the x-space compared to the rest of the data points. However, it closely aligns with the regression line that passes through the remaining sample points. Such a point is referred to as a "leverage point," and it possesses certain unique characteristics:

- Point A does not exert influence on the estimates of the regression coefficients themselves.
- Nevertheless, it does have an impact on various model summary statistics, including R-squared (R^2) and the standard errors of regression coefficients.

Now, let's shift our focus to point B in the Fig. 10.20:

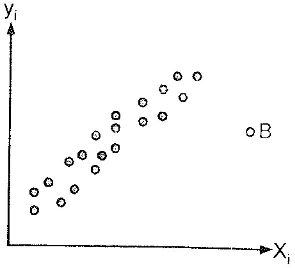


Fig. 10.20 Influence point

Point *B*, in contrast, exhibits a moderately unusual x-coordinate, and its corresponding *y*-value is also atypical. This categorizes it as an 'influence point':

- An influence point significantly affects the model coefficients.
- It exerts a pulling effect on the direction of the regression model.

At times, a small subset of data can wield a disproportionately strong influence on the model coefficients and its overall characteristics. In the most extreme scenario, the parameter estimates might end up relying more on this influential subset of data points than on most of the dataset. This situation is less than ideal, as a regression model should ideally represent the entire spectrum of sample observations, rather than being predominantly shaped by just a few data points. Therefore, it becomes essential to identify these influential data points and assess their impact on the model.

- If these influential points are determined to be outliers or erroneous values, it may be necessary to exclude them from the sample.
- However, if there is nothing inherently problematic with these points but they significantly influence the model's properties, it becomes imperative to understand precisely how they are affecting the regression model currently in use.

10.8.1 Leverage

The positioning of data points within the x-space has a notable impact on various aspects of the model, including parameter estimates, standard errors, predicted values, and summary statistics. An essential tool for identifying influential observations is the hat matrix denoted as H, which is calculated as $H = X(X^T X)^{-1} X^T$.

Given that

$$\boxed{\begin{aligned} Var(\hat{y}) &= \sigma^2 H \\ Var(e) &= \sigma^2 (I - H) \end{aligned}}$$

Where $\hat{y}$ is fitted value and e is residual. The elements h_{ii} of the hat matrix *H* provide valuable insights. Specifically, these h_{ii} elements can be interpreted as quantifying the level of leverage exerted by the i^{th} observation y_i on the i^{th} fitted value $\hat{y}_i$.

The i^{th} diagonal element of *H* is

$$\boxed{h_{ii} = x_i' (X^T X)^{-1} x_i}$$

where x_i represents the i^{th} row of X-matrix. The diagonal of the hat matrix is a standardized measure of the distance of i^{th} observation from the centre (or centroid) of the x-space. Thus, the large hat diagonals reveal observations that are potentially influential because they are situated at a considerable distance within the x-space, distinct from the bulk of the sample data points.

Average size of hat diagonal ($\bar{h}$) is

$$\bar{h} = \frac{\Sigma h_{ii}}{n} = \frac{rank(H)}{n} = \frac{rank(X)}{n} = \frac{trace(H)}{n} = \frac{k}{n}$$

When $h_{ii} > 2\bar{h} = \frac{2k}{n}$, it indicates that the data point is sufficiently distant from the remaining data, and it can be categorized as a leverage point.

It's important to exercise caution when setting the cutoff value at $2k/n$, and careful consideration should be given to evaluating the magnitudes of both k and n. There may be scenarios where $2k/n > 1$, and in such cases, this cutoff criterion does not apply.

Regarding point A in Fig. 10.19:

- It will exhibit a notably large hat diagonal and can unquestionably be classified as a leverage point.
- However, it will not impact the regression coefficients, given that it lies on the same line that traverse through the other observations.

When assessing the influence of observations, the hat diagonal primarily considers the spatial location of observations within the x-space. Therefore, it's advisable to also examine the studentized residual or R-student in conjunction with the diagonal element h_{ii}. Observations characterized by both a substantial hat diagonal and large residuals are likely to be influential.

10.8.2 Measures of Influence

(1) Cook's D-statistics:

In cases where the dataset is limited in size, removing data points can significantly impact the quality of the fit and the validity of statistical conclusions.

When evaluating the influence of data points, it is advantageous to consider two key factors:

- The spatial location of the data point within the x-space.
- The behaviour and characteristics of the response variable.

The Cook's distance statistics denoted as, Cook's D-statistic is a measure of the distance between the least squares estimate based on all n observations and the estimate obtained by deleting the i^{th} point

$$D_i = \frac{e_i^2}{k - MSS_{res}} \times \frac{h_{ii}}{(1-h_{ii})^2} = \frac{r_i^2}{k} \times \left(\frac{h_{ii}}{1-h_{ii}}\right)$$

It considers both the leverages and the residual of the observation. High Cook's distance indicates that the observation has a large effect on the estimated coefficients when it's deleted.

If the Cook's D is higher than 1.0, or $2 \times \sqrt{(k/n)}$, the observation is highly likely to be influential. The Cook's distance is useful for identifying influential data points as it measure the change in regression estimate if the observation is deleted.

(2) DFFITS (Difference in Fits):

Cook's distance measure is a deletion diagnostic, *i.e.*, it measures the influence of i^{th} observation if it is removed from the sample.

DFFITS is a statistical measure used in regression analysis to assess the influence of individual data points on the regression model's predictions. It quantifies how much the fitted values (predicted values) would change if a specific data point were omitted from the model. DFFITS stands for 'Difference in Fits'.

$$DFFITS_i = \frac{(\hat{y}_i - \hat{y}_{(i)})}{\sqrt{s_{(i)}^2 h_{ii}}}$$

In the formula, $\hat{y}_{(i)}$ represents the predicted value for the i^{th} observation when it is excluded from the model. The denominator in the formula serves as a standardization factor because it accounts for the $Var(\hat{y}_i) = \sigma^2 h_{ii}$, Essentially, this *DFFITS* value signifies how many standard deviations the fitted valuey $\hat{y}_i$ would change if the i^{th} observation were to be removed from the model.

An observation is considered influential when the absolute value of its DFFITS value surpasses the threshold of:

$$\boxed{2\sqrt{\frac{k+2}{n-k-2}}}$$

Here, as always, n represents the total number of observations, and k represents the count of predictor terms (excluding the intercept) in the regression model. It's important to note that this is not a rigid rule but rather a general guideline. Different authors may use slightly different guidelines. Consequently, a more subjective criterion might be preferred, such as deeming a data point influential when the absolute magnitude of its *DFFITS* value conspicuously deviates from the pattern exhibited by the other *DFFITS* values. This subjective assessment acknowledges the inherent subjectivity in identifying outliers, which are, by nature, subjective phenomena.

3. DFBETAS (Difference in Betas)

DFBETAS is a statistical measure which quantifies how much the estimated coefficients (betas) would change if a particular data point were excluded from the model. DFBETAS stands for 'Difference in Betas'.

Mathematically, *DFBETAS* for the j^{th} predictor variable and i^{th} observation is calculated as follows:

$$\boxed{DFBETAS_i = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{s_{(i)}^2 C_{ii}}}}$$

Where:

- $\hat{\beta}_j$ represents the estimated coefficient for the j^{th} predictor when all data points are considered.
- $\hat{\beta}_{j(i)}$ represents the estimated coefficient for the j^{th} predictor when the i^{th} data point is excluded from the model.
- C_{ii} is i^{th} diagonal element of $(X^T X)^{-1}$ when all data points are considered.

Interpreting DFBETAS:

A large absolute value of *DFBETAS* suggests that excluding the corresponding data point significantly affects the estimated coefficient $\hat{\beta}_j$.

Observations with $|DFBETAS_i| > 2/n$ are often considered influential for the corresponding predictor variable.

DFBETAS is a valuable diagnostic tool in regression analysis, helping identify influential data points that may have a substantial impact on the regression coefficients. It is used alongside other diagnostic measures to assess model assumptions and the overall quality of the regression model.

4. Covariance Ratio (COVRATIO)

The covariance ratio statistic measures the change in the determinant of the covariance matrix of the estimates by deleting the i^thobservation

$$\boxed{COVRATIO_i = \frac{| X_{(i)}^T X_{(i)})^{-1} s_{(i)}^2 |}{| (X^T X)^{-1} MSS_{res} |} \quad i = 1, 2, \ldots, n}$$

- When $COVRATIO_i > 1$, it indicates that including the i^{th} observation improves the precision of estimation.
- Conversely, if $COVRATIO_i < 1$, adding the i^{th} observation may reduce the precision of estimation.

COVRATIO is a valuable tool for identifying influential observations in regression analysis. High or low *COVRATIO* values suggest that certain data points have a notable impact on the model's performance and parameter estimates. Typically, observations with *COVRATIO* values significantly deviating from 1 are considered influential and warrant further investigation.

Cutoff values for *COVRATIO* are not easy to obtain. It is suggested that if

$$\boxed{\begin{array}{l} COVRATIO_i > 1+\dfrac{3k}{n} \\ \\ COVRATIO_i < 1-\dfrac{3k}{n} \end{array}}$$

Or

Then i^{th} point should be considered influential. The lower bound is only appropriate when $n > 3k$. These cutoffs are only recommended for large samples.

10.9 VARIABLE SELECTION

The entire process of conducting a regression analysis depends on the choice of explanatory predictor variables included in the model. It is a fundamental principle in regression analysis that only the correct and essential explanatory variables should be part of the model. In practice, after confirming the appropriate functional form for the model, analysts often have a pool of potential explanatory variables that could potentially influence the process or experiment. However, it's not common to include all of these candidate variables in the regression model; instead, a subset of explanatory variables is carefully selected from this pool. This process of determining the appropriate subset of explanatory variables for regression is known as 'variable selection'.

When it comes to selecting a subset of explanatory variables, two primary options exist:

Inclusion of as Many Explanatory Variables as Possible: One approach is to include as many explanatory variables as feasible to create a model that closely mirrors reality.

Simplicity with Fewer Explanatory Variables: Another approach is to keep the model as simple as possible by incorporating only a limited number of explanatory variables.

Both of these approaches have their implications. When numerous variables are included in the model, they can influence the prediction of the study variable y. Conversely, when only a small number of variables are used, it reduces the predictive variance of $\hat{y}$. Additionally, collecting data on a larger number of observations can be costlier and more time-consuming. Striking a balance between these considerations is essential to select the 'best regression equation'.

The challenge of variable selection assumes that the functional form of the explanatory variables (*e.g.*, $x^2, \dfrac{1}{x}, \log x$, etc.) is already known, and there are no outliers or influential observations in the data. Various statistical tools, including residual analysis, identification of outliers or high-leverage data points, and assessments of model adequacy, are linked to the variable selection process. These steps are often carried out iteratively. Initially, a strategy for variable selection is chosen, and the model is constructed with the selected variables. The fitted model is then scrutinized for the correct functional form, the presence of outliers, influential observations etc. Based on the outcome, the model is re-examined, and the selection of variable is reviewed again. Several iterations may be required before the final adequate model is decided.

There can be two types of incorrect model specifications.

1. Omission/exclusion of relevant variables.
2. Inclusion of irrelevant variables.

10.9.1 Omission/Exclusion of Relevant Variables

To maintain model simplicity, the analyst may opt to exclude certain explanatory variables that could be theoretically important. This decision may be influenced by several factors. For instance, it could be challenging to quantitatively measure variables such as taste or intelligence. Additionally, there might be difficulties in obtaining accurate observations for variables like income.

Suppose there are k candidate explanatory variables, out of which r variables are chosen to be included in the model, and $(k-r)$ variables are designated for exclusion. We can partition the design matrix X and the coefficient vector β as follows:

$$\underset{n\times k}{X} = \begin{pmatrix} X_1 & X_2 \\ n\times r & n\times(k-r) \end{pmatrix} \text{ and } \underset{n\times k}{\beta} = \begin{pmatrix} \beta_1 & \beta_2 \\ r\times 1 & (k-r)\times 1 \end{pmatrix}$$

The model $y = X\beta + \varepsilon, E(\varepsilon) = 0, \text{Var}(\varepsilon) = \sigma^2 I$ can be expressed as:

$$\boxed{y = X_1\beta_1 + X_2\beta_2 + \varepsilon}$$

which is called a full model or true model.

After dropping the [illegible] explanatory variable in the model, the new model is

k-r

$$y = X_1\beta_1 + \delta.$$

which is called a **mis specified model** or **false model**.

Applying OLS (ordinary least square) to the false model, the OLSE of β_1.

$$\hat{\beta}_{1F} = \left(X_1^T X_1\right)^{-1} X_1^T y$$

Estimation Error　$$\hat{\beta}_{1F} = \left(X_1^T X_1\right)^{-1} X_1^T (X_1\beta_1 + X_2\beta_2 + \varepsilon)$$

$$\hat{\beta}_{1F} = (X_1^T X_1)^{-1} X_1^T X_1 \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T \varepsilon$$

Taking expectation on both sides

$$E\left[\hat{\beta}_{1F}\right] = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 E[\beta_2] + (X_1^T X_1)^{-1} X_1^T E[\varepsilon]$$

$$\boxed{E\left[\hat{\beta}_{1F}\right] = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 \neq \beta_1}$$

So, unless $X_1^T X_2 = 0$, $\hat{\beta}_{1F}$ is biased. The bias can be huge. The bias is vanished when X_1 and X_2 are orthogonal or uncorrelated.

$$\boxed{E\left[\hat{\beta}_{1F}\right] - \beta_1 = \theta}$$

Where $\theta = (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$ and is which is a linear function of β_2, i.e., the coefficients of excluded variables.

Mean Square Error (MSE)

$$E\left[\left(\hat{\beta}_{1F} - \beta_1\right)\left(\hat{\beta}_{1F} - \beta_1)^T\right)\right] = E\left[\left(\theta + (X_1^T X_1)^{-1} X_1^T \varepsilon\right)\left(\theta + (X_1^T X_1)^{-1} X_1^T \varepsilon\right)^T\right]$$

$$E\left[\theta\theta^T + (X_1^T X_1)^{-1} X_1^T \varepsilon\theta^T + \theta\varepsilon^T X_1 (X_1^T X_1)^{-1} + (X_1^T X_1)^{-1} X_1^T \varepsilon\varepsilon^T X_1 (X_1^T X_1)^{-1}\right]$$

$$= \boxed{\theta\theta^T + \sigma^2 (X_1^T X_1)^{-1}}$$

The mean squared error (MSE) of a mis-specified model is higher than that of the full model, resulting in a decrease in efficiency.

Variance of $\hat{\beta}_{1F}$

$$\text{Var}\ (\hat{\beta}_{1F}) = E\left[\hat{\beta}_{1F} - E\left[\hat{\beta}_{1F}\right]\right]^2$$

$$= E\left[\hat{\beta}_{1F} - (\beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2)\right]^2$$

$$E\left[(X_1^T X_1)^{-1} X_1^T (X_1\beta_1 + X_2\beta_2 + \varepsilon) - (\beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2))\right]^2$$

$$= E\left[(X_1^T X_1)^{-1} X_1^T \varepsilon\varepsilon^T X_1 (X_1^T X_1)^{-1}\right] = \boxed{\sigma^2 (X_1^T X_1)^{-1}}$$

We get a smaller variance when we omit X_2.

Key points of Omitting Relevant Variables

Omitting relevant variables from a regression analysis can indeed lead to inconsistency and bias in estimating the effects of variables, even though it may reduce the variance of the estimator. This is a fundamental trade-off in regression analysis, and it's known as the bias-variance trade-off.

Here's a breakdown of this trade-off:

Bias: When relevant variables are omitted from the model, it can introduce bias in the estimates of the coefficients. This bias occurs because the model cannot account for the true relationships between the omitted variables and the dependent variable. As a result, the estimated coefficients for the included variables may be systematically over or underestimated.

Inconsistency: Omitting relevant variables can also make the parameter estimates inconsistent. Inconsistent estimates mean that as you collect more data, the estimated coefficients do not converge to the true population values. This inconsistency can lead to unreliable and misleading results.

Variance: On the other hand, omitting variables may reduce the variance of the parameter estimates. This can make the estimates less sensitive to random fluctuations in the data. However, this reduction in variance comes at the cost of increased bias and inconsistency.

To strike a balance, it's essential to carefully consider which variables should be included in the model based on domain knowledge, theoretical understanding, and empirical evidence. Techniques like stepwise regression or regularization methods (e.g., Ridge or Lasso regression) can also be used to handle variable selection while attempting to mitigate the bias-variance trade-off.

In summary, while omitting relevant variables may reduce variance, it introduces bias and inconsistency into the analysis. Striking the right balance between bias and variance is a crucial aspect of building reliable regression models.

10.9.2 Inclusion of Irrelevant Variables

Another common scenario involves adding variables to the regression equation that lack economic relevance. Researchers may be inclined to include variables solely based on their statistical significance, even if some of these variables hold no economic relevance. In the classical linear regression model, the inclusion of an economically irrelevant variable does not introduce bias into the least squares estimates but does increase their variance.

Let the true model is

$$y = X\beta + \varepsilon, E(\varepsilon) = 0, \text{Var}(\varepsilon) = \sigma^2 I$$

which comprise k explanatory variable. Suppose now r additional explanatory variables are added to the model and the resulting model becomes

$$\boxed{y = X\beta + Z\gamma + \delta}$$

where Z is a $n \times r$ matrix of n observations on each of the r explanatory variables and γ is $r \times 1$ vector of regression coefficient associated with Z and δ is disturbance term. This model is termed as a false model.

Applying OLS false model, we get

$$\begin{bmatrix} \hat{\beta}_F \\ \hat{Y}_F \end{bmatrix} = \begin{bmatrix} X^T X & X^T X \\ Z^T X & Z^T Z \end{bmatrix}^{-1} \begin{bmatrix} X^T y \\ Z^T y \end{bmatrix}$$

$$\begin{vmatrix} X^T y \\ Z^T y \end{vmatrix} - \begin{vmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z \end{vmatrix} \begin{vmatrix} \hat{\beta}_F \\ \hat{Y}_F \end{vmatrix}$$

where $\hat{\beta}_F$ and $\hat{\gamma}_F$ are the OLSEs of β and γ respectively.

After solving, we get

$$\hat{\beta}_F = (X^T \bar{H}_z X)^{-1} X^T \bar{H}_z y$$

where $$\bar{H}_z = [1 - Z(Z^T Z)^{-1} Z^T]$$

Estimation Error:

$$E[\hat{\beta}_F] = E[(X^T H_z X)^{-1} X^T H_z y] - E[(X^T H_z X)^{-1} X^T H_z (X\beta + \varepsilon)] = \boxed{\beta}$$

Therefore, $\hat{\beta}_F$ is unbiased even when some irrelevant variables are added to the model.

Variance of $\hat{\boldsymbol{\beta}}_F$:

$$\text{Var}(\hat{\beta}_F) = E\left[\hat{\beta}_F - E\left[\hat{\beta}_F\right]\right]^2$$

$$E\left[\hat{\beta}_F - \beta\right]^2 = E[(X^T H_z X)^{-1} X^T H_z (X\beta + \varepsilon) - \beta]^2$$

$$= E[(X^T H_z X)^{-1} X^T H_z X\beta\beta^T X^T H_z{}^T X (X^T H_z X)^{-1} + 2(X^T H_z X)$$
$$+ (X^T H_z X)^{-1} X^T H_z X\varepsilon\varepsilon^T X^T H_z{}^T X (X^T H_z X)^{-1} + \beta\beta^T - 2(X^T H_z X)^{-1} X^T H_z X \beta\beta^T$$
$$- 2(X^T H_z X)^{-1} X^T H_z \varepsilon$$

$$E\left[\hat{\beta}_F - \beta\right]^2 = E[(X^T H_z X)^{-1} X^T H_z \varepsilon\varepsilon^T H_z{}^T X (X^T H_z X)^{-1}] = \boxed{\sigma^2 (X^T H_z X)^{-1}}$$

The variance of the model increases when irrelevant variables are included.

Key points of Inclusion of irrelevant variables

Including explanatory variables that are not truly relevant to the model can indeed have unintended consequences. Let's delve into this issue further:

Reduction in Degrees of Freedom: When irrelevant variables are included in the model, they consume degrees of freedom without contributing significantly to explaining the variation in the dependent variable. This reduction in degrees of freedom can affect the reliability of statistical tests and model evaluation.

Coefficient of Determination (R-squared): The *R*-squared value measures the proportion of variance in the dependent variable explained by the independent variables. Including irrelevant variables tends to increase R-squared, making the model appear better at explaining the data than it is. This can be misleading and overstate the model's goodness of fit.

Overfitting: Including too many irrelevant variables can lead to overfitting, where the model fits the noise in the data rather than the true underlying patterns. Overfit models perform well on the training data but generalize poorly to new, unseen data.

Model Complexity: More variables often lead to a more complex model. Complex models may be harder to interpret and may not be suitable for making predictions or drawing meaningful conclusions.

10.10 MODEL BUILDING

A question arises once subsets of candidate variables have been selected for the model: how can we assess which subset results in a better regression model? Various criteria have been proposed in the literature to evaluate and compare these subset regression models.

10.10.1 Coefficient of Determination

The coefficient of determination represents the square of the multiple correlation coefficient between the study variable, denoted as "y," and the set of explanatory variables: $X_1, X_2, \dots, X_p$. This is expressed as R^2. It's important to note that for all cases, $i = 1, 2, \dots, n$, $X_1 = 1$, indicating the necessity of an intercept term in the model. Without an intercept term, the coefficient of determination cannot be employed. Consequently, there will be a subset of $(p - 1)$ explanatory variables and one intercept term in the notation r_p^2.

The coefficient of determination based on these variables is calculated as:

$$r_p^2 = \frac{SSR_p}{SST} = 1 - \frac{SSE_p}{SST}$$

where SSR_p and SSE_p are the sum of squares due to regression and residuals, respectively in a subset model based on $(p - 1)$ explanatory variables.

Given the availability of k explanatory variables and the selection of only $(p - 1)$ out of them, there $\binom{k}{p-1}$ possible choices of subsets, each resulting in a unique subset model. It's important to note that the coefficient of determination $\left(r_p^2\right)$ tends to increase as the number of variables, p, increases.

To determine the optimal value of p, follow these steps:

1. Choose an initial value of p, fit the model, and calculate $r_p{}^2$).
2. Add one variable to the model, fit the updated model, and obtain r_{p+1}^2.
3. Evaluate whether the addition of the extra variable results in a substantial increase in r_{p+1}^2. Specifically, check if $r_{p+1}^2 - r_p^2$ is relatively small. If the increase in r_p^2 is minimal, you may consider stopping and selecting the value of p for your subset regression.
4. If, on the other hand, the increase in r_p^2 is substantial, continue adding variables until you reach a point where the increment in r_p^2 becomes small or an additional variable no longer produces a significant change in the value of r_p^2.

To identify this optimal value of p, create a plot of r_p^2 versus p. For example, the curve will exhibit behaviour as depicted in the figure 10.20.

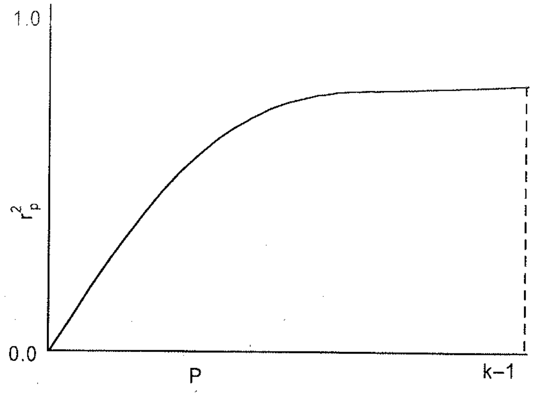


Fig. 10.21 Effect of increasing p on r_p^2

This method allows you to determine the appropriate number of variables to include in your subset regression model, achieving a balance between model complexity and explanatory power.

10.10.2 Coefficient of Determination

The adjusted coefficient of determination has certain advantages over the usual coefficient of determination. The adjusted coefficient of determination $r_{adj}^2(p)$ based on p-term model is

$$r_{adj}^2(p) = 1 - \frac{n-1}{n-p}(1 - r_p^2)$$

An advantage of $r_{adj}^2(p)$ is that it does not necessarily increase as p increases. If there are r more explanatory variables which are added to term model, then if and only if the corresponding F statistic for inclusion of that variable is greater than one, then only it increases.

10.10.3 Residual Mean Square

A model is considered to have a better fit when the residuals are minimized, as reflected in the sum of squares due to residuals, SSE_p. A model with a smaller SSE_p is preferred. Building on this concept, the residual mean square MSE_p, based on a subset regression model with p variables, is defined as:

$$MSE_p = \frac{SSE_p}{n-p}$$

Therefore, MSE_p can serve as a criterion for model selection, just like SSE_p. The value of SSE_p tends to decrease as p increases. Similarly, as p increases, MSE_p initially decreases, reaches a stable point, and may eventually increase if the model is unable to compensate for the loss of one degree of freedom due to the factor $(n - p)$.

When MSE_p is plotted against p, the curve exhibits behavior as depicted in the following figure.

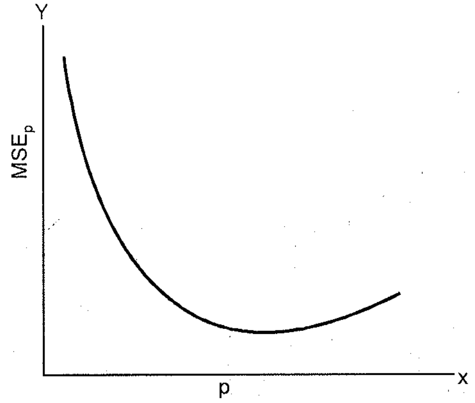


Fig. 10.22 Effect of increasing *p* on MSE$_p$

So

- Plot MSE_p versus p.
- Choose p corresponding to the minimum value of MSE_p.
- Choose p corresponding to which MSE_p is approximately equal to MSE based on the full model.
- Choose p near the point where the smallest value MSE_p turns upward.

Such minimum value of MSE_p will produce $r^2_{adj}(p)$ with maximum value. So, the two criteria, viz, minimum MSE_p and maximum $r^2_{adj}(p)$ are equivalent.

10.10.4 Mallow's C$_p$ Statistics

Mallow's C_p criterion is derived from the mean squared error of a fitted value. For a model with p parameters, we define C_p as:

$$C_p = \frac{SSE_p}{\sigma^2} + 2p - n$$

When evaluating various subset models, the models with the smallest value of C_p are regarded as superior to those with higher C_p values. Therefore, a lower C_p is the preferred choice.

If the model includes all important explanatory variables, then $E[SSE_p] = (n - p)\sigma^2$, and as a result, $E(C_p)$ becomes:

$$E[C_p] = \frac{(n-p)\sigma^2}{\sigma^2} + 2p - n = p$$

Conversely, if the model excludes important variables, $E[SSE_p] > (n-p)\sigma^2$, leading to $E(C_p) > p$. Therefore, the preference is to opt for a model with a C_p value close to p.

Additionally, it can be demonstrated that C_p serves as an estimator of the mean square error of prediction, given by:

$$\frac{1}{n}\sum_{i=1}^{n}[Var(\hat{y}_i) + \{bias\ (\hat{y}_i)\}^2]$$

This implies that minimizing C_p is desirable. Hence, you should choose either:

- The model that minimizes C_p, or
- A model with C_p close to p, with p being small.

The plot of C_p versus p for each regression equation will be a straight line passing through the origin and looks like as follows:

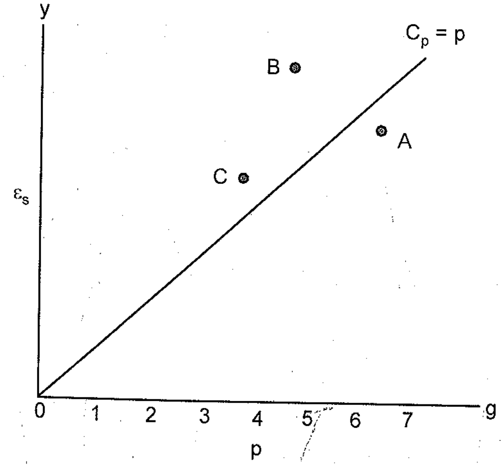


Fig. 10.23 Plot of C_p versus p

Points that exhibit smaller bias tend to cluster near the regression line, while those with substantial bias typically deviate above the line. For instance, consider point A, which displays minimal bias and consequently resides closer to line A. In contrast, points B and C exhibit significant bias, positioning them above the line. Furthermore, point C is situated higher than point A, representing a model with reduced overall error. In certain situations, it might be advantageous to tolerate some bias in the regression equation to achieve a reduction in the average prediction error.

It's important to highlight that Mallow's C_p, where $C_p = p$, employs an unbiased estimator of σ^2 with the assumption that the full model possesses minimal bias. However, if the full model incorporates non-significant explanatory variables with zero regression coefficients, the same unbiased estimator of σ^2 can result in an overestimation of σ^2, leading to smaller values for C_p. Consequently, the effectiveness of C_p hinges on the selection of a suitable estimator for σ^2.

10.10.5 Akaike's Information Criterion (AIC)

The Akaike's Information Criterion, often denoted as AIC, is a statistical measure expressed as:

$$AIC_p = n \ln \frac{SSR_p}{n} + 2p$$

In this equation,

$SSR_p = y^T X_1 (X_1^T X_1)^{-1} X_1^T y$ is based on the subset model derived from full model

$$y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon = X\beta + \varepsilon, E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2 I$$

AIC is formally defined as:

AIC = – 2 (maximized log likelihood) + 2(number of parameters).

In the linear regression model with $\varepsilon \sim N(0, \sigma^2 I)$, the likelihood function is:

$$L(y, \beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{(y - X\beta)^T (y - X\beta)}{2\sigma^2}\right]$$

and log-likelihood of $L(y, \beta, \sigma^2)$ is:

$$\ln L(y, \beta, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\left[\frac{(y - X\beta)^T (y - X\beta)}{\sigma^2}\right]$$

The log likelihood is maximized as:

$$\tilde{\beta} = (X^T X)^{-1} X^T y$$

$$\tilde{\sigma}^2 = \frac{n-p}{n}\tilde{\sigma}^2$$

where $\tilde{\beta}$ is maximum likelihood estimate of β which is same as OLSE $\tilde{\sigma}^2$ is maximum likelihood estimate of σ^2 and $\tilde{\sigma}^2$ is OLSE of σ^2.

Therefore,;

$$AIC = -2 \ln L\left(y, \tilde{\beta}, \tilde{\sigma}^2\right) + 2p = n \ln \frac{SSR}{n} + 2p + n[\ln(2\pi) + 1)$$

where $SSR = y^T [1 - X(X^T X)^{-1} X^T] y$

AIC is a valuable tool for model selection and comparison, taking into consideration both the model's likelihood and the number of parameters involved. Lower AIC values indicate models that strike a better balance between goodness of fit and complexity.

10.10.6 Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC), also known as the Schwarz Criterion, is a statistical criterion used for model selection and model comparison in statistical modelling and machine learning. BIC is designed to balance the trade-off between model fit and model complexity, similar to Akaike's information Criterion (AIC). However, BIC tends to penalize models with more parameters more heavily than AIC.

The formula for BIC is as follows:

$$BIC_p = -2 \ln L + (k - n) \ln n$$

Where:

- $ln(L)$ is the maximum value of the likelihood function of the model.
- k is the number of parameters in the model.
- n is the sample size (number of observations).

The BIC has the following characteristics:

It includes a penalty for the number of parameters, which discourages overfitting.

BIC tends to favour simpler models when the sample size is small.

It is particularly useful when there is a need to select the best-fitting model from a set of candidates.

In practice, a lower BIC value indicates a better model in terms of balancing fit and complexity, making it a useful criterion for selecting the most appropriate model among competing alternatives.

10.10.7 PRESS Statistic

Just as the residuals and residual sum of squares serve as criteria for selecting subset models, similarly, the PRESS residuals and prediction sum of squares can also be employed as a foundation for subset model selection. Both the regular residuals and PRESS residuals possess unique characteristics that find application in regression modelling.

The PRESS statistic based on a subset model with p explanatory variable is given by,

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} \left(\frac{e_i}{(1 - h_{ii})} \right)^2$$

where h^{ii} is the i^{th} element in $H = X(X^T X)^{-1} X^T$. This criterion is used on similar lines as in the case of SSR_p. A subset regression model with a smaller value of $PRESS_p$ is preferable.

10.10.8 Partial F-statistics

Suppose the overall regression model as tested by the Analysis of Variance table is significant. We know that not all of the ? parameters are zero. But we may still be able to delete several variables. We can carry out the Subset test based on the extra sum of squares principle.

The partial F-statistic is used to test the hypothesis about a sub vector of the regression coefficient.

Consider the model $\underbrace{y}_{n\times 1} = \underbrace{X}_{n\times p}\ \underbrace{\beta}_{p\times 1} + \underbrace{\varepsilon}_{n\times 1}$

Where

where $p = k+1$ which includes an intercept term and k explanatory variables. Suppose a subset of $r < k$ explanatory variables is to be obtained which contribute significantly to the regression model. So, partition

$$X = [X_1 \quad X_2], \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

where X_1 and X_2 are matrices of order $n \times (p-r)$ and $n \times r$, respectively; β_1 and β_2 are the vectors of order $(p-1) \times 1$ and $r \times 1$, respectively.

The objective is to test the null hypothesis

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Then

$y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$ is the full model and application of least squares gives the OLSE of β as:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

The corresponding sum of squares due to regression with p degrees of freedom is

$$SSR = \hat{\beta}^T X^T y - n\bar{y}^2$$

and the sum of squares due to residuals with $(n-p)$ degrees of freedom is

$$SSE = y^T y - \hat{\beta}^T X^T y$$

and $$MSE = \frac{SSE}{n-p}$$

Similarly, for the reduced model $y = X_1\beta_1 + \delta, E(\delta) = 0$ has

$$SSR(\text{red}) = \hat{\beta}_1^T X_1^T y - n\bar{y}^2$$

$$SSE(\text{red}) = y^T y - \hat{\beta}_1^T X_1^T y$$

The sum of squares of regression due to β_2 given that β_1 in already in the model can be found by

$$SSR(\beta_2/\beta_1) = SSR - SSR(\text{red}) = \hat{\beta}^T X^T y - \hat{\beta}_1^T X_1^T y$$

The term $SSR(\beta_2/\beta_1)$ is called as the extra sum of squares due to β_2 and has degrees of freedom $p - (p-r) = r$. It is a measure of regression sum of squares that results from adding the explanatory variables $X_{k-r+1}, \ldots, X_k$ in the model when the model has already $X_1, \ldots X_r$ explanatory variables.

The null hypothesis H_0: $\beta_2 = 0$ can be tested using the statistic

$$F_0 = \frac{SSR(\beta_2/\beta_1)/r}{MSE}$$

which follows F-distribution with r and $(n-p)$ degrees of freedom under H_0. The decision rule is to reject H_0 whenever $F_0 > F_\alpha(r, n-p)$. This is known as the partial F-test. It measures the contribution of explanatory variables in X_2 given that the other explanatory variables in X_1 are already in the model.

10.11 COMPUTATIONAL TECHNIQUES FOR VARIABLE SELECTION

To choose a subset model, various techniques relying on computational methods and algorithms are available. These techniques primarily revolve around two fundamental concepts: including all potential explanatory variables or opting for a stepwise selection of explanatory variables.

10.11.1 Use all Possible Explanatory Variables

This methodology is based on the following steps:

Fit a model with one explanatory variable.

Fit a model with two explanatory variables.

Fit a model with three explanatory variables.

and so on.

Choose a suitable criterion for model selection and evaluate each of the fitted regression equation with the selection criterion. The total number of models to be fitted sharply rises with an increase in k. So such models can be evaluated using a model selection criterion with the help of an efficient computation algorithm on computers.

10.11.2 Stepwise Regression Techniques

This methodology is based on choosing the explanatory variables in the subset model in steps which can be either adding one variable at a time or deleting one variable at times. Based on this, there are three procedures.

- Forward selection,
- backward elimination and
- stepwise regression.

These procedures are basically computer-intensive procedures and are executed using the software.

Forward selection procedure

This methodology assumes that there is no explanatory variable in the model except an intercept term. It adds variables one by one and tests the fitted model at each step using some suitable criterion. It has the following steps.

- Consider only intercept term and insert one variable at a time.
- Calculate the simple correlations of x_i's ($i = 1, 2, \ldots, k$) with y
- Choose x_i which has the largest correlation with y.
- Suppose x_1 is the variable which has the highest correlation with y. Since F-statistic given by

$$F_0 = \frac{n-k}{k-1} \cdot \frac{r^2}{1-r^2}$$

So x_1 will produce the largest value of F_0 in testing the significance of a regression.

- Choose a prespecified value of F value, say F_{IN} (F to enter).
- If $F > F_{IN}$, then accept x_1 and so x_1 enters into the model.
- Adjust the effect of x_1 on y and re-compute the correlations of remaining x_i's with y and obtain the partial correlations as follows.
    - Fit the regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ and obtain the residuals.
    - Fit the regression of x_1 on other candidate explanatory variables as $\hat{x}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} x_1$, $j = 2, 3, \ldots, k$ and obtain the residuals.
    - Find the simple correlation between the two residuals.
    - This gives the partial correlations.
- Choose x_i with the second-largest correlation with y, *i.e.*, the variable with the highest value of partial correlation with y.
- Suppose this variable is x_2. Then the largest partial F-statistic is
- $F = \dfrac{SSR(x_2 | x_1)}{MSE(x_1, x_2)}$
- It $F > F_{IN}$ then x_2 enters into the model.
- These steps are repeated. At each step, the partial correlations are computed, and explanatory variable corresponding to the highest partial correlation with y is chosen to be added into the model. Equivalently, the partial F-statistics are calculated, and the largest F-statistic given the other explanatory variables in the model are chosen. The corresponding explanatory variable is added into the model if partial F-statistic exceeds F_{IN}.
- Continue with such selection as long as either at a particular step, the partial F-statistic does not exceed F_{IN} or when the least explanatory variable is added to the model.

Backward Elimination Procedure:

This methodology is contrary to the forward selection procedure. The forward selection procedure starts with no explanatory variable in the model and keeps on adding one variable at a time until a suitable model is obtained. The backward elimination methodology begins with all explanatory variables and keeps on deleting one variable at a time until a suitable model is obtained. It is based on the following steps:

- Consider all k explanatory variables and fit the model.
- Compute partial F-statistic for each explanatory variables as if it were the last variable to enter in the model.
- Choose a preselected value F_{OUT} (F-to-remove).
- Compare the smallest of the partial F-statistics with F_{OUT}. If it is less than F_{OUT}, then remove the corresponding explanatory variable from the model.
- The model will have now $(k - 1)$ explanatory variables.
- Fit the model with these $(k - 1)$ explanatory variables, compute the partial F-statistic for the new model and compare it with F_{OUT}. If it is less than F_{OUT}, then remove the corresponding variable from the model.
- Repeat this procedure.
- Stop the procedure when the smallest partial F-statistic exceeds F_{OUT}.

Stepwise regression procedure: Stepwise regression is a technique that combines aspects of both forward selection and backward elimination procedures. It is essentially an adaptation of the forward selection method and encompasses the following steps:

- Begin by considering all the explanatory variables that have been included in the model at the previous step.
- Add a new variable to the model and perform regression using their partial F-statistics.
- An explanatory variable that was added in an earlier step may now become insignificant due to its relationship with the currently present explanatory variables in the model.
- If the partial F-statistic for an explanatory variable is smaller than a predefined F_{OUT} value, that variable is removed from the model.
- The stepwise regression process involves the use of two threshold values: F_{IN} and F_{OUT}. In some cases, F_{IN} may be equal to F_{OUT}, or F_{IN} may be greater than F_{OUT} The choice of using F_{IN} greater than F_{OUT} makes it relatively more challenging to introduce a new explanatory variable into the model compared to removing one.

General Comments

1. None of the methods among the forward selection, backward elimination or stepwise regression guarantees the best subset model.
2. The order in which the explanatory variables enter or leave the models does not indicate the order of importance of the explanatory variable.

PROBLEMS

1. What do you mean by positive skewed distribution?
2. How would you define variance stabilizing transformations and what purpose do they serve?
3. What is your comprehension of measures of influence? Elaborate on two specific measures of influence.
4. Explain how bias and variance are affected by omitting the variables from the model.
5. Explain how the model complexity is related with the number of variables in the model.
6. Explain the residual mean square of a model?
7. What are the various Computational techniques for variable selection?
8. What are the common assumptions taken over a data set of a model.
9. What do you understand by the term "explanatory variables" of a model.
10. Define residual analysis and describe how it is used for model diagnostics. Provide examples of situations where residuals might indicate problems with the model.

CHAPTER

11

LOGISTIC REGRESSION MODEL AND POISSON REGRESSION MODEL

Chapter Outline

- Logictic Regression Model • Linear Predictor and Link Functions • Maximum Likelihood Estimation of Parameters • Maximum Likelihood Estimation of Parameters • Poisson Regression Model • Linear Predictor and Link Functions • Maximum Likelihood Estimation of Parameters • Hypothesis Testing of Poisson Regression Model

11.1 LOGISTIC REGRESSION MODEL

In certain regression scenarios, the response variable y is characterized by only two potential outcomes. For instance, it could be high blood pressure versus low blood pressure, the development of esophageal cancer versus not developing it, whether a crime will be solved or remain unsolved, or whether a bee specimen belongs to the 'killer' (Africanized) bee category or the domestic honeybee category. In these instances, we can represent the outcome y as either 0 or 1, and our objective is to predict this outcome, or its probability based on one or more predictor variables (x's).

In the linear regression model $y = X\beta + \varepsilon$, we encounter two distinct types of variables: explanatory variables, denoted as $X_1, X_2, ..., X_k$, and the study variable, denoted as y. These variables can be measured along a continuous scale or represented as indicator variables.

When the explanatory variables are qualitative in nature, their values are typically expressed using indicator variables, leading to the application of dummy variable models.

Conversely, when the study variable is qualitative, it assumes values that can be effectively represented using an indicator variable with only two possible outcomes: 0 and 1. In such instances, logistic regression becomes the preferred statistical technique. For example, y may represent qualitative states such as success or failure, yes or no, like or dislike, which can be conveniently encoded as binary values 0 and 1.

Consider the model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

$$y_i = x'_i\beta + \varepsilon_i \quad \text{for} \quad i = 1, 2, \ldots, n$$

Where

$$x'_i = [1\ x_{1i}\ x_{2i} \ldots x_{ki}] \quad \text{and} \quad \beta = [\beta_0\ \beta_1 \ldots \beta_k]$$

The study variable assumes one of two values: 0 when $y = 0$ or 1 when $y = 1$. We can assume that y follows a Bernoulli distribution with a parameter π, which means its probability distribution is as follows:

$$y_i = \begin{cases} 1\ with\ P(y_i = 1) = \pi_i \\ 0\ with\ P(y_i = 0) = 1 - \pi_i \end{cases}$$

Assuming, $E[\varepsilon_i] = 0$

$$E[y_i] = 1 \times \pi_i + 0 \times (1 - \pi_i) = \pi_i$$

From the model $y_i = x'_i\,\beta + \varepsilon_i$, we have:

$$E[y_i] = E[x'_i\,\beta + \varepsilon_i] = \pi_i$$

$$E[y_i] = P(y_i = 1)$$

Thus response function $E[y_i]$ is simply the probability $P(y_i = 1)$.

Note that

Model $\varepsilon_i = y_i - x'_i\,\beta$, hence

- When $y_i = 1$ then $\varepsilon_i = 1 - x'_i\,\beta$
- $y_i = 0$ then $\varepsilon_i = -x'_i\beta$

Recall our previous assumption that ε_i followed a normal distribution when y was not an indicator variable. However, when y takes on the role of an indicator variable, ε_i is constrained to only two possible values, and thus, it cannot be presumed to follow a normal distribution.

In the typical regression model, the errors exhibit homoskedasticity, meaning that $Var(\varepsilon_i) = \sigma^2$ and consequently, $Var(y_i) = \sigma^2$. However, when y is treated as an indicator variable, this relationship changes.

$$\begin{aligned} Var(y_i) &= E[y_i - E[y_i]]^2 \\ Var(y_i) &= (1 - \pi_i)^2\,\pi_i + (0 - \pi_i)^2\,(1 - \pi_i) \\ &= \pi_i\,(1 - \pi_i)\,[1 - \pi_i + \pi_i] = \pi_i\,(1 - \pi_i) \\ &= E(y_i)\,[1 - E(y_i)] \\ &= \sigma^2_{y_i} \end{aligned}$$

Consequently, the variance $Var(y_i)$ now depends on the value of y_i and becomes a function of the mean of y_i. Therefore, the usual least-squares estimators $\hat{\beta}$ will not be optimal. Furthermore, since $E(y_i) = \pi_i$, where π_i is the probability, it follows that $0 \leq \pi_i \leq 1$. Therefore, there exists a constraint on $E(y_i)$, which must fall within the range of 0 to 1. This constraint significantly limits the options for selecting the linear response function. It becomes impossible to construct a model in which the predicted values extend beyond the interval of 0 and 1.

When y is a binary variable, empirical evidence indicates that the function $E(y)$, which maps from the entire real number line to the interval [0,1], typically exhibits a sigmoidal shape. This shape is characterized by a nonlinear S-curve, similar to the following:

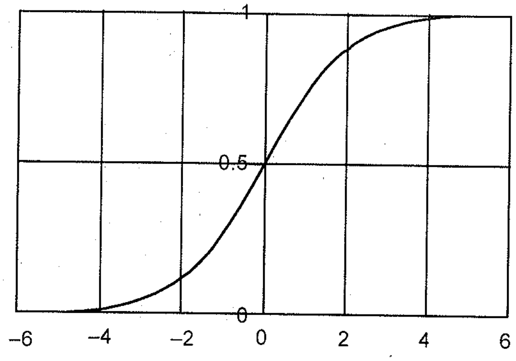


Fig. 11.1 Logistic Function

An appropriate selection for $E(y)$ can be the cumulative distribution function of a random variable. Specifically, the logistic distribution, characterized by its simplified logistic function as its cumulative distribution function, provides a suitable link and can be expressed as:

$$
\begin{aligned}
E(y) &= \frac{\exp(y)}{1+\exp(y)} \\
&= \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \\
&= \frac{1}{1+\exp(-x_i'\beta)}
\end{aligned}
$$

11.2 LINEAR PREDICTOR AND LINK FUNCTIONS

The systematic component within $E(y)$ is referred to as the linear predictor and is symbolized as:

$$\eta_i = \sum_j \beta_j x_{ji} = x_i'\beta$$

The link function in generalized linear model relates the linear predictor η_i to the mean response μ_i.

Thus, $$g(\mu_i) = \eta_i$$

or $$\mu_i = g^{-1}(\eta_i)$$

In typical linear models that rely on a normally distributed study variable, the link function "$g(\mu_i) = \mu_i$" is employed, and it's known as an identity link. This link function effectively maps the range of μ_i onto the entire real number line, offering a robust empirical approximation and allowing for meaningful interpretations in practical applications.

In logistic regression, the link function is specified as:

$$\eta = \ln\frac{\pi}{1-\pi}$$

This transformation is called as the logit transformation of probability μ and $\frac{\pi}{1-\pi}$ is called as odds. The link η is also called as log-odds. This link function is obtained as follows:

$$\pi = \frac{1}{1+\exp(-\eta)}$$

or $$\pi\,[1 + \exp(-\eta)] = 1$$

$$\exp(-\eta) = \frac{1-\pi}{\pi}$$

or $$\mu = \ln\frac{\pi}{1-\pi}$$

Please note that, like the logit function, there exist other functions that share the same shape as the logistic function. These functions can also be adjusted through π. Among the two popular functions in this category are the probit transformation and the complementary log-log transformation.

The probit transformation is based on the adjustment of π using the cumulative distribution function of the normal distribution and serves as the foundation for the probit regression model.

The complementary log-log transformation of π is: $\ln[-\ln(1-\pi)]$

11.3 MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS

Consider the general form of the logistic regression model

$$\text{Model } y_i = E(y_i) + \varepsilon_i'$$

where y_i are independent Bernoulli random variable with a parameter π_i with

$$E(y_i) = \pi_i = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$$

The probability density function of y_i is

$$f_i\,(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, i = 1, 2, \ldots, n, \; y_i = 0 \text{ or } 1$$

The likelihood function is:

$$L(y_1, y_2, \ldots, y_n, \beta_1, \beta_2, \ldots, \beta_k) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

Taking logarithm both sides, we have:

$$\ln L(y_1, y_2, \ldots, y_n; \beta_1, \beta_2, \ldots, \beta_k) = \ln L(\beta) = \sum_{i=1}^{n}\left[\ln \pi_i^{y_i} + \ln(1-\pi_i)^{1-y_i}\right]$$

$$= \sum_{i=1}^{n}[y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)]$$

$$= \sum_{i=1}^{n}\left[y_i \ln \frac{\pi_i}{(1-\pi_i)} + \ln(1-\pi_i)\right]$$

As

$$\pi_i = \frac{\exp\,(x_i'\,\beta)}{1+\exp\,(x_i'\,\beta)}$$

$$1 - \pi_i = \frac{1}{1+\exp\,(x_i'\,\beta)}$$

And

$$\frac{\pi_i}{(1-\pi_i)} = \exp\,(x_i'\,\beta)$$

$$\ln \frac{\pi_i}{(1-\pi_i)} = (x_i'\,\beta)$$

$$\ln L(\beta) = \sum_{i=1}^{n}[y_i \ln x_i'\,\beta - \ln(1+\exp(x_i'\,\beta))]$$

11.4 HYPOTHESIS TESTING OF LOGISTIC REGRESSION MODEL

The hypothesis testing for the parameters within the logistic regression model relies on asymptotic theory. It involves a large sample test employing the likelihood ratio test based on a statistic known as deviance.

A saturated model, which consists of exactly p parameters and perfectly fits the sample data, serves as a reference point. The statistic used to compare the log-likelihoods between the fitted model and the saturated model is referred to as the model deviance. This model deviance is defined as:

$$\lambda(\beta) = 2 \ln L(\text{saturated model}) - 2 \ln L(\hat{\beta})$$

Here, $\ln L(\cdot)$, represents the log-likelihood, and $\hat{\beta}$ represents the maximum likelihood estimate of β.

In the case of the logistic regression model, $y_i = 0$ or 1 and π_i's are completely unrestricted. So, the likelihood will be maximum at $\pi_i = y_i$, and the maximum value of L (saturated model) is

$$Maximum\ L(\text{saturated model}) = 1$$

$$ln(Maximum\ L(saturated\ model) = 0$$

Let $\hat{\beta}$ be the maximum likelihood estimation of β, then log-likelihood is maximum at $\beta = \hat{\beta}$ and

$$ln\, L(\hat{\beta}) = \sum_{i=1}^{n} [y_i \, ln\, x_i' \hat{\beta} - ln(1 + \exp(x_i'\hat{\beta})]$$

$$\geq \ln L \text{ (saturated model)}$$

Assuming the correctness of the logistic regression function, the large sample distribution of the likelihood ratio test statistic $\lambda(\beta)$ is approximately distributed as $\chi^2(n-p)$ when n is sufficiently large.

A high value of $\lambda(\beta)$ indicates that the model is not a good fit, while a low value of $\lambda(\beta)$ suggests that the model is well-fitted and performs comparably to the saturated model. It's important to note that, in general, the fitted model will have fewer parameters than the saturated model, which incorporates all parameters. Therefore, at a significance level of α %, we can assess whether the model is a suitable fit or not.

11.5 POISSON REGRESSION MODEL

The traditional regression model is built upon the assumption that the random errors conform to a normal distribution, and consequently, the study variable follows a normal distribution as well. However, when the study variable is dichotomous and can only take binary values, specifically 0 and 1, logistic regression is employed. In this case, the study variable adheres to a Bernoulli distribution.

Similarly, we encounter scenarios where the study variable represents a count, typically associated with relatively rare events. For instance, it might be the count of patients with a rare disease, with one or more explanatory variables like age, haemoglobin level, and blood sugar. In another example, the study variable could represent the number of defects in car engines produced by a reputable car manufacturer, and this count depends on one or more explanatory variables.

In such situations, the assumption of a normal or Bernoulli distribution for the study variable is not suitable. Instead, the Poisson distribution provides a more appropriate description. Hence, we assume that the study variable, denoted as y, is a count variable and follows a Poisson distribution with a parameter $\lambda > 0$ as follows:

$$f(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

Note that the mean and variance of a Poisson random variable are same and related as:

$$E(y) = \lambda, \quad Var(y) = \lambda$$

Based on a sample $y_1, y_2, \ldots, y_n$ we can write

$$E(y_i) = \lambda$$

and express the Poisson regression model as

$$y_i = E[y_i] + \varepsilon_i, \qquad i = 1, 2, \ldots, n$$

where ε_i's are disturbance terms.

11.6 LINEAR PREDICTOR AND LINK FUNCTIONS

We can define a link function g that relates to the mean of study variable to a linear predictor as:

$$\begin{aligned} g(\lambda_i) &= \eta_i \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \\ &= x'_i \beta \\ \lambda_i &= g^{-1}(\eta_i) = g^{-1}(x'_i\beta) \end{aligned}$$

The identity link function is

$$g(\lambda_i) = x'_i \beta$$

The log-link function is

$$g(\lambda_i) = \ln(\lambda_i) = x'_i \beta$$

$$\Rightarrow \qquad \lambda_i = g^{-1}(x'_i \beta) = \exp(x'_i \beta)$$

Note: That in identity link function, the predicted values of y can be negative but in log-link function, the predicted values of y are nonnegative.

11.7 MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS

We use the method of maximum likelihood estimation to estimate the parameters of the poisson regression model. The likelihood function is based on Poisson distribution with parameter λ and then β's are estimated through the link function.

$$L(y, \lambda) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

$$\ln L(y, \lambda) = \sum_{i=1}^{n} y_i \ln \lambda_i - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \ln y_i!$$

Upon selecting the appropriate link function, you can maximize the log-likelihood function using numerical optimization techniques with a given dataset. Let $\widehat{\beta}$ represent the resulting maximum likelihood estimator of β. Consequently, the Poisson regression model that has been fitted is:

$$\hat{y}_i = g^{-1}\left(x_i'\hat{\beta}\right)$$

In case of identity link,

$$\hat{y}_i = g^{-1}\left(x_i'\hat{\beta}\right) = x_i'\hat{\beta}$$

In the case of log-link

$$\hat{y}_i = g^{-1}\left(x_i'\hat{\beta}\right) = \exp\left(x_i'\hat{\beta}\right)$$

11.8 HYPOTHESIS TESTING OF POISSON REGRESSION MODEL

The test of hypothesis in the case of Poisson regression model is similar to the case of the logistic regression model. It is constructed as model deviance which is based on a large sample test using the likelihood ratio test statistic. The model deviance is defined as

$$\lambda(\beta) = 2 \ln L(\text{saturated model}) - 2 \ln L(\hat{\beta})$$

The saturated model, which encompasses all p parameters of the model, is designed to perfectly fit the data. The statistic $\lambda(\beta)$ follows an approximate $\chi^2(n-p)$ distribution when n is sufficiently large. A high value of $\lambda(\beta)$ suggests that the model is not a good fit for the given data, while small values of $\lambda(\beta)$ indicate that the model is well-fitted and performs well as the saturated model.

If $\lambda(\beta) \leq \chi^2_{n-p}(\alpha)$, then the fitted model is considered adequate. Conversely, if $\lambda(\beta) > \chi^2_{n-p}(\alpha)$, then the fitted model is deemed inadequate at the α% significance level.

PROBLEMS

1. What do you mean by decision boundary in classification problems?
2. How do you make a prediction using Logistic regression Model.
3. What is the key difference between linear regression and logistic regression?
4. What is sigmoid function?
5. What does the SoftMax function entail and in what contexts is it employed?
6. What are the advantages of logistic regression?
7. What are the disadvantages of logistic regression?
8. How is the performance of a Logistic Regression Model typically evaluated?
9. What are the coefficients in logistic regression, and how do they influence a prediction?
10. How is the logistic regression model employed to depict the correlation between features and the likelihood of the event taking place?
11. What is the Poisson regression model used for?
12. What type of probability distribution is commonly associated with the Poisson Regression Model?
13. How does the Poisson Regression Model express the relationship between the mean and the predictor variables?
14. What is the key assumption of the Poisson regression model regarding the variance of the response variable?
15. When does overdispersion happen in Poisson regression, and how can we fix it?
16. Given a logistic regression model with coefficients: $b_0 = -3$ and $b_1 = 0.8$, calculate the probability of an event $Y = 1$ for a data point with a feature value of $X = 5$. Use the logistic function $P(Y = 1) = \dfrac{1}{1 + e^{-(b_0 + b_1 X)}}$.
17. Suppose you have a Logistic Regression Model with the following coefficients: $b_0 = 1.5$, $b_1 = -0.2$, and $b_2 = 0.03$. If the feature values for a specific data point are $X_1 = 8$ and $X_2 = 12$, calculate the probability of the event $Y = 1$ using the logistic regression equation

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 X + b_2 X_2)}}$$

18. Suppose you have a Poisson regression model for predicting the number of accidents Y based on the number of hours of exposure to traffic X. The model equation is $\log(\mu) = 2 + 0.5X$, where μ is the mean of the Poisson distribution. If a location has 10 hours of exposure, calculate the predicted mean number of accidents.
19. Given a Poisson Regression Model with the equation $\log(\mu) = -1 + 0.2\,X_1 - 0.1\,X_2$, where X_1 represents the number of advertisements and X_2 represents the time spent on a website, calculate the predicted mean number of customer purchases? for a scenario where $X_1 = 8$ and $X_2 = 20$.

□□□

CHAPTER

12

DATA ANALYTICS

Chapter Outline

• Real Analysis and Data Science • Basic set Theory • Norms and Metric • Open Set and Closed Set • Sequences of Numbers • Compact Sets • Completeness in Metric Spaces • Connected Metric Space

12.1 REAL ANALYSIS AND DATA SCIENCE

Real analysis is a branch of mathematics that focuses on the study of real numbers and the properties and behaviour of functions of real variables. It is a fundamental area of mathematical analysis that provides a rigorous foundation for many other branches of mathematics, including calculus, probability theory, and functional analysis. Real Analysis is important for data science for the following reasons:

1. **Foundation for Statistics:** Real analysis provides a rigorous foundation for statistical concepts and methods used in data science. Probability distributions, sampling, hypothesis testing, and statistical inferences are all deeply rooted in real analysis.
2. **Understanding Limits and Convergence:** Concepts like limits and convergence are a fundamental in part of real analysis. In data science, especially when dealing with algorithms and optimization, understanding these concepts is crucial. For instance, in machine learning, convergence of optimization algorithms is a key consideration.
3. **Continuous Mathematics:** Real analysis deals with continuous functions and spaces. Many mathematical models in data science, such as those used in machine learning and optimization, involve continuous variables and functions. Real analysis provides the mathematical tools which help in rigorously analyzing and understand these continuous structures.

4. **Functional Analysis:** Functional analysis is a branch of mathematics rooted in real analysis, which helps in understanding spaces of functions and their properties. This is relevant in data science when working with function spaces, such as spaces of probability distributions or function spaces in functional data analysis.
5. **Optimization:** Optimization is a key aspect of data science, particularly in machine learning where models are trained to optimize certain objectives. Real analysis is crucial in studying optimization problems, convergence of optimization algorithms, and the properties of optimal solutions.
6. **Probability Theory:** Probability theory, which acts as a lease for statistics and data science, heavily relies on concepts from real analysis. The development of probability spaces, random variables, and the analysis of probability distributions often involve real analysis techniques.
7. **Modelluing and Simulation:** Real analysis provides the mathematical tools to model and simulate real-world phenomena. This is essential in data science when developing mathematical models to represent complex systems while simulating scenarios for predictive modelluing.
8. **Mathematical Rigor:** Real analysis promotes mathematical rigor and precision. Data science involves making informed decisions based on data, and having a solid mathematical foundation which ensures that the analysis and conclusions are accurate and reliable.

In summary, real analysis serves as a foundational framework for various mathematical concepts and techniques that are widely employed in data science. It enhances the understanding of mathematical structures, facilitates rigorous analysis, and provides necessary tools for modelluing, optimization, and statistical reasoning in the field of data science. Before exploring the intricacies of real analysis, it's essential to provide a brief overview of set theory. This basic understanding is crucial for grasping the key concepts that real analysis introduces.

12.2 BASIC SET THEORY

Set theory is a branch of mathematical logic that studies sets, which are collections of objects or elements. It serves as the foundation for much of modern mathematics, providing a formal language for describing mathematical structures and relationships.

Sets

A set is a well-defined collection of distinct objects, considered as an element in its own right. The objects within a set are called elements. Sets are denoted by curly braces, and elements are listed inside. For example, if A is a set of the first three natural numbers, it can be denoted as $A=\{1, 2, 3\}$.

Membership

The symbol $\in$ is used to indicate that an element belongs to a set. For example, $1 \in A$ is read as '1 belongs to the set A'.

Equality

Two sets are equal if and only if they have exactly the same elements. The order in which elements are listed does not matter.

Subset

If every element of set A is also an element of set B, then A is a subset of B, denoted as $A \subseteq B$. If A is a subset of B but A is not equal to B, then A is a proper subset of B, denoted as $A \subset B$.

Universal Set and Complement

The universal set, denoted by U, is the set that contains all elements under consideration. The complement of a set A with respect to the universal set U is the set of elements in U that are not in A, denoted as A' or A^c.

Intersection and Union

The union of sets A and B, denoted as $A \cup B$, is the set of elements in either A or B or both.

$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$

The intersection of sets A and B, denoted as $A \cap B$, is the set of elements common to both A and B.

$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$

Complement and Set Difference

The complement of a set A with respect to a universal set U, denoted by A' or A^c, is the set of all elements in U that are not in A.

$$A' = \{x \in U : x \notin A\}$$

The difference of sets A and B, denoted by $A - B$ or $A \setminus B$, is the set of all elements in A that are not in B.

$$A - B = \{x \in A : x \notin B\}$$

Cartesian Product

The Cartesian product of sets A and B, denoted as $A \times B$, is the set of all ordered pairs (a, b) where $a \in A$ and $b \in B$. This concept generalizes to the Cartesian product of more than two sets.

$$A \times B = \{(a, b) : a \in A \text{ and } b \in B\}$$

Power Set

The power set of a set A, denoted as $P(A)$ or $\mathbf{2}^A$, is the set of all possible subsets of A, including the empty set and A itself.

$$P(A) = \{X : X \subseteq A\}$$

Set theory is often formulated using axioms to establish a foundation for the rest of mathematics. Zermelo-Fraenkel set theory with the Axiom of Choice (ZFC) is a widely accepted framework.

Understanding these basic concepts of set theory is important to work with mathematical structures and proofs in various branches of mathematics, including algebra, analysis, and topology. It provides a language for expressing mathematical ideas precisely and rigorously.

12.3 NORMS AND METRIC

A normed space is characterized as a vector space equipped with a norm, providing a meaningful measure of vector length. On the other hand, a metric space is defined as a set furnished with a metric, offering significance to the distance between any two points within the set. It's noteworthy that there is always a metric associated with a norm. 'Normed' spaces constitute a subset within the category of metric spaces, and in turn, metric spaces comprise a subset within the realm of topological spaces.

The **Euclidean distance** (or **metric**) between two points $x, y \in \mathbb{R}^n$ is defined as

$$\left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}.$$

In mathematics, a norm is a function that assigns a positive size or length to vectors in a vector space. Formally, let V be a vector space over the field of real or complex numbers. A norm on V is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies the following properties for all vectors $x, y \in V$ and all scalars $\alpha \in \mathbb{R}$:

1. Non-negativity: $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$.
2. Scalar multiplication: $\|ax\| = |a| \cdot \|x\|$.
3. Triangle inequality: $\|x+y\| \leq \|x\| + \|y\|$.

A common example of a norm is the Euclidean norm (or 2-norm) in $\mathbb{R}^n$, given by

$$\|v\| = \sqrt{x^2 + y^2 + z^2}$$

A metric space is a mathematical structure that provides a way to quantify the concept of distance between points in a set. It is a fundamental concept in the field of topology and analysis.

Definition: A metric space is a set X equipped with a function $d : X \times X \to \mathbb{R}$ called a metric, which satisfies the following properties for all points x, y, z in X:

1. Non-negativity: $d(x, y) \geq 0$ (Distance is always non-negative).
2. Identity of indiscernibles: $d(x, y) = 0$ if and only if $x = y$ (Distance is zero only for identical points).
3. Symmetry: $d(x, y) = d(y, x)$ (Distance between x and y is the same as between y and x).
4. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$ (The shortest distance between two points is a straight line).

Example 12.1 ℓ^1 metric

Define d_1: $\mathbb{R}^n \times \mathbb{R}^n \to [0, \infty]$, such that

$$d_1(x, y) = \sum_{i=1}^{n} | x_i - y_i | .$$

This is called the ℓ^1 **metric**.

Example 12.2 ℓ^p **metric**

Define d_p: $\mathbb{R}^n \times \mathbb{R}^n \to [0, \infty]$, such that

$$d_p(x, y) = \left(| x_i - y_i |^p \right)^{1/p}$$

It is called the ℓ^p metric.

12.3.1 Components of a Metric Space

Set X: This is the collection of points for which we want to define distances.

Metric *d*: The metric is a function that assigns a non-negative real number (distance) to pairs of points in X. It captures the intuitive notion of how 'far apart' points are from each other.

Example12.3 Euclidean Space: In $\mathbb{R}$, the Euclidean distance is a metric. For two points $(x_1, y_1, \ldots, z_1)$ and $(x_2, y_2, \ldots, z_2)$, the distance is given by the square root of the sum of squared differences.

Example 12.4 Discrete Metric: In any set X, the discrete metric is defined as $d(x, y) = 1$ if $x \neq y$ and $d(x, y) = 0$ if $x = y$. This metric captures whether two points are the same or different.

12.3.2 Importance of Metric Space

1. *Topology:* Metric spaces are foundational in topology, a branch of mathematics that studies the properties of spaces that remain unchanged under continuous deformations.
2. *Analysis:* In analysis, the concept of convergence of sequences and continuity of functions is often defined using metrics.
3. *Geometry:* Metric spaces provide a generalization of the notion of distance, allowing mathematicians to study spaces with different geometrical properties.

12.4 OPEN SET AND CLOSED SET

12.4.1 Open Set

A set $U \subset X$ is called open, if for each $x \in U$ there exists an $\in > 0$ such that the interval $(x - \in, x + \in)$ is contained in U. Such an interval is often called an $\in$-neighbourhood of x, or simply a neighbourhood of x.

Alternatively, it can also be stated that in a metric space X with a metric d, a subset $U \subset X$ is an open set if, for every point $x \in U$, there exists an open ball centered at x that is entirely contained in U.

Example12.5 Consider the open interval (a, b) in $\mathbb{R}$ with the standard Euclidean metric. For any point x in (a, b), there exists an open interval around x that is entirely contained in (a, b). Therefore, (a, b) is an open set in the metric space $\mathbb{R}$.

12.4.2 Closed Set

A set $F \subset X$ is called closed if the complement of F, $X \setminus F$, is open.

Example12.6 Consider the closed interval $[a, b]$ in $\mathbb{R}$ with the standard Euclidean metric. The complement of $[a, b]$ is $(-\infty, a) \cup (b, \infty)$, which is open. Therefore, $[a, b]$ is a closed set.

Example12.7 Which of the intervals $(-4, 4)$, $[3, 8]$, $(-4, 5]$, $(0, \infty)$ and $[0, \infty)$ are open, closed, both, or neither?

- The interval $(-4, 4)$ is open, because if x is any number in $(-4, 4)$, then $-4 < x < 4$ or equivalently, $-4 - x < 0 < 4 - x$. Now let $\in = \min(4 + x, 4 - x)$. Then $\in > 0$, and the interval $\left(x - \frac{\in}{2}, x + \frac{\in}{2}\right)$ is contained in $(-4, 4)$.
- The interval $[3, 8]$ is closed, because its complement consists of the two open sets $(-\infty, 3)$ and $(8, \infty)$.
- The interval $(-4, 5]$ is neither open nor closed. It is not open, because the point $x = 5$ is contained in the set, but every neighbourhood of that point is not contained

inside the set. It is not closed, because its complement $(-\infty, -4]$ and $(5, \infty)$ is not open at $x = -4$.

- The interval $(0, \infty)$ and also the interval $(-\infty, 0)$ are both open, because every point in the set contains a neighborhood contained inside the set.
- The interval $[0, \infty)$ and also the interval $(-\infty, 0]$ are both closed, because their complements are the open sets mentioned above.

Example 12.8 The sets $\mathbb{R}$ (the whole real line) and ϕ (the empty set) are open and closed simultaneously.

- Take any point x contained in $\mathbb{R}$. Clearly, any neighbourhood of x consists of real numbers, and hence is also contained in $\mathbb{R}$. Therefore, $\mathbb{R}$ is open.
- Take the empty set. Since it does not contain any point, we do not have to check whether any neighbourhoods are contained in the empty set. It is therefore open vacuously.
- Since the complement of the real line is the empty set, and the empty set is open, the real line is also closed.
- Since the complement of the empty set is the real line, and the real line is open, the empty set is also closed.

12.4.3 Unions of Open Sets, Intersections of Closed Sets

- Every union of open sets is again open.
- Every intersection of closed sets is again closed.
- Every finite intersection of open sets is again open.
- Every finite union of closed sets is again closed.

12.4.4 Example of Open and Closed Set

The entire metric space X and the empty set ϕ are both open and closed. These sets are sometimes referred to as clopen sets.

12.4.5 Example of Half-Open Interval

The half-open interval $[0,1)$ in $\mathbb{R}$ is open but not closed. Its complement is $(-\infty, 0) \cup [1, \infty)$, which is not open.

Understanding open and closed sets in a metric space is fundamental for analysing the topology of the space and for defining continuity, convergence, and other key concepts in analysis.

12.5 SEQUENCES OF NUMBERS

12.5.1 Sequences

Sequences are, basically, many countable numbers arranged in an order that may or may not exhibit certain patterns. Here is the formal definition of a sequence:

A **sequence** of real numbers is a function $f : \boldsymbol{N} \rightarrow \boldsymbol{R}$. In other words, a sequence can be written as $f(1)$, $f(2)$, $f(3)$,..... Generally, such sequencesis denoted by the symbol $\{a_j\}_{j=1}^{\infty}$, wherein $a_j = f(j)$.

A sequence $\{a_j\}_{j=1}^{\infty}$, is ***bounded*** if the underlying function is bounded. That is, if there exists a $B \in \mathbb{R}$ such that

$$|a\_j| \leq B \text{ for all } j \in \boldsymbol{N}$$

In other words, the sequence $\{a_j\}_{j=1}^{\infty}$ is bounded whenever the set $\{a_j : j \in \boldsymbol{N}\}$ is bounded. We similarly define the words ***bounded below*** and ***bounded above***.

When we need to give a concrete sequence, we often give each term as a formula in terms of n. For example, $\{1/N\}_{n=1}^{\infty}$ stands for the sequence $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots$ This sequence is a bounded sequence. On the other hand, the sequence $\{n\}_{n=1}^{\infty}$ stands for 1, 2, 3, 4,..., and this sequence is not bounded.

Another example of a sequence is the so-called *constant sequence*. That is a sequence $\{c\}_{n=1}^{\infty} = c, c, c, c....$ consisting of a single constant $c \in \mathbb{R}$ repeating indefinitely.

12.5.2 Convergence

A sequence $\{a_j\}_{j=1}^{\infty}$ of real (or complex) numbers is said to **converge** to a real (or complex) number c if for every $\in > 0$ there is an integer $N > 0$ such that if $j > N$ then

$$|a_j - c| < \in$$

The number c is called the limit of the sequence $\{a_j\}_{j=1}^{\infty}$ and we sometimes write $a_j \rightarrow c$.

If a sequence $\left\{a_j\right\}_{j=1}^{\infty}$ does not converge, then we say that it **diverges**.

Convergent sequences, in other words, exhibit the behaviour that they get closer and closer to a particular number.

Example12.9 Consider the sequence $\left\{a_j\right\}_{j=1}^{\infty}$. It converges to zero. Prove it.

Solution. $\left\{a_j\right\}_{j=1}^{\infty} = \{1, 1/2, 1/3, 1/4, \ldots\}$, which seems to indicate that the terms are getting closer and closer to zero. According to the definition of convergence, we need to show that no matter which $\in > 0$ one chooses, the sequence will eventually become smaller than this number. To be precise: take any $\in > 0$. Then there exists a positive integer N such that $\frac{1}{N} < \in$. Therefore, for any $j > N$ we have:

$$| 1/j - 0 | = | 1/j | < 1/N < \in$$

whenever $j > N$.

12.5.3 Cauchy Sequence

Let $\left\{a_j\right\}_{j=1}^{\infty}$ be a sequence of real (or complex) numbers. We say that the sequence satisfies the Cauchy criterion (or simply is Cauchy) if for each $\in > 0$ there is an integer $N > 0$ such that if $j, k > N$ then

$$|a_j - a_k| < \in$$

This definition states precisely what it means for the elements of a sequence to get closer together, and to stay close together.

12.6 COMPACT SETS

In math, especially in general topology, compactness is a property that tries to generalize the idea of a closed and bounded subset in Euclidean space. The main concept is that a compact space doesn't have any 'holes' or 'missing points'; it includes all the points where sequences converge. For instance, the open interval (0, 1) wouldn't be compact because it doesn't include the endpoints 0 and 1, while the closed interval [0, 1] would be compact as it includes these endpoints. Similarly, the set of rational numbers $\mathbb{Q}$ isn't compact because it has infinitely many 'holes' corresponding to irrational numbers. The set of real numbers R isn't compact either because it doesn't include the two infinite values, positive and negative infinity. However, the extended real number line would be compact because it includes both infinities.

12.6.1 Sequential Compactness

Imagine you have a bunch of points in a space, and you want to know if every time you pick an endless list of these points, some of them will get super close to a particular point in that space. If that's the case, we say the space is 'sequentially compact'.

There's a special rule for the regular space we're used to, like the points on a graph. It's called the Bolzano-Weierstrass theorem, and it says that if you want a part of this space to be compact in this special way, that part must be both closed and bounded.

Let's use an example to understand this. If you take a bunch of points in the closed unit interval [0, 1], some of those points will get really close to some number between 0 and 1. But if you try the same thing with the open unit interval (0, 1), where you don't include the endpoints, it won't work. The points might get close to 0 or 1, but they won't get close to any point inside the interval.

Even though parts of our regular space can be compact, the whole space itself isn't compact because it goes on forever. Think about the real number line — if you take the sequence 0, 1, 2, 3, ..., there's no bunch of points that get close to any real number.

A subset *K* of a topological space is said to be compact if, intuitively, every open cover of *K* has a finite subcover. In more formal terms:

Definition: *A* set *K* is compact if, for every open cover $\{U_\alpha\}$ of *K* (where each U_α is an open set), there exists a finite subset of this cover such that *K* is still covered.

12.6.2 Key Properties of Compact Sets

1. **Closed and Bounded (in Euclidean Spaces)**

   In Euclidean spaces (like $\mathbb{R}^n$), a subset *K* is compact if and only if it is closed and bounded. This is known as the **Heine-Borel Theorem**.

2. **Finite Sub-cover**

   The essential property of compactness is the existence of a finite sub-cover for every open cover. No matter how 'spread out' the open sets are, we can always find a finite number of them that still cover the entire set.

3. **Compactness Implies Closed and Bounded (in Euclidean Spaces)**

   In Euclidean spaces, compactness implies closed and bounded property. However, in more general topological spaces, compact sets might not exhibit these properties.

4. **Compactness Preserved under Continuous Maps**

   If $f : X \to Y$ is a continuous map between topological spaces and *X* is compact, then the image *f*(*X*) is compact in *Y*.

12.6.3 Examples of Compact Sets

1. Closed and Bounded Intervals (in $\mathbb{R}$):

   The closed and bounded intervals $[a, b]$ in $\mathbb{R}$ are compact sets.

2. Finite Sets:

   Any finite set in any topological space is compact.

12.6.4 Importance of Compact Sets

1. **Analysis and Topology:** Compact sets play a crucial role in analysis and topology. They provide a way to generalize concepts like continuity, convergence, and completeness.
2. **Existence of Extrema:** In calculus, the extreme value theorem relies on the compactness of closed and bounded intervals to ensure the existence of global extrema.
3. **Functional Analysis:** In functional analysis, a branch of mathematics that studies vector spaces of functions, compactness is a key concept in various theorems and definitions.

Understanding compact sets is fundamental in advanced mathematical studies, and their properties have applications in various areas of mathematics, including analysis, topology, and functional analysis.

12.7 COMPLETENESS IN METRIC SPACES

We've established that every compact subset within a metric space is both closed and bounded. However, it's important to highlight thatall closed and bounded set are not considered compact. In this section, our focus shifts towards a comprehensive characterization of compact sets: a set is deemed compact if and only if it is both 'compelete' and'totally bounded.

To define completeness, we need to again define Cauchy sequences. A sequence has a limit if its terms get close to some point. A sequence is Cauchy when its terms "get close to each other". A metric space is complete if every Cauchy sequence has a limit. The idea is that a Cauchy sequence looks like it should be converging to something, and a space is complete if there is always something there for it to converge to.

Definition: A metric space is said to be complete if every Cauchy sequence in that space converges to a limit that is also in the space. In other words, a metric space is complete if every Cauchy sequence in that space converges to a limit within the space itself. Formally, for every Cauchy sequence x_n in the metric space, there exists a point x in the space such that x_n converges to x asn approaches infinity.

Completeness is a desirable property because it ensures that the metric space has 'no missing points' in a certain sense. Every Cauchy sequence, which intuitively represents a sequence of points getting arbitrarily close to each other, has a limit within the space itself.

12.7.1 Examples of Complete Metric Space

1. Consider the real numbers with the standard Euclidean metric. $\mathbb{R}$ is a complete metric space because every Cauchy sequence of real numbers converges to a real limit, and that limit is also a real number.
2. In contrast, the rational numbers $\mathbb{Q}$ are not complete under the same metric. There exist Cauchy sequences of rational numbers that do not converge to a rational limit within $\mathbb{Q}$. The irrational numbers fill the 'gaps' and to make $\mathbb{Q}$ complete, one needs to extend it to $\mathbb{R}$.
3. The Euclidean space $\mathbb{R}^n$ is complete. Cauchy sequences of vectors in $\mathbb{R}^n$ converge to a limit vector in $\mathbb{R}^n$.
4. Closed intervals, such as $[a, b]$ in $\mathbb{R}$ are complete metric spaces. Any Cauchy sequence of real numbers within the interval converges to a limit within the same interval.
5. Many Banach spaces are complete metric spaces. Banach spaces are normed vector spaces where the metric is induced by the norm. Examples include L_p spaces for $1 \le p \le \infty$.

Theorem 1: Let (M, d) be a complete metric space, and let $\boldsymbol{X}$ be a subset of $\boldsymbol{M}$. The subspace (X, d) is a complete metric space if and only if $\boldsymbol{X}$ is a closed subset of $\boldsymbol{M}$.

Proof : If (x_n) is a Cauchy sequence in (X, d) then it is also a Cauchy sequence in (M, d). Since (M, d) is complete then (x_n) converges. If $\boldsymbol{X}$ is closed then the limit of (x_n) is in $\boldsymbol{X}$. Hence, (X, d) is a complete metric space. Now suppose that (X, d) is a complete metric space and let (x_n) be a sequence in $\boldsymbol{X}$ that converges to $x \in \boldsymbol{M}$. Then (x_n) is a Cauchy sequence in $\boldsymbol{M}$ and thus also Cauchy in $\boldsymbol{X}$. Since $\boldsymbol{X}$ is complete then $x \in X$. Hence, $\boldsymbol{X}$ is closed.

Theorem 2 : Let (K, d) be a compact metric space. Then, (K, d) is a complete metric space.

Proof: Suppose (K, d) is compact. Let $\{x_n\}_{n=1}^{\infty}$ be a Cauchy sequence in K. We show that $\{x_n\}_{n=1}^{\infty}$ converges, thus proving K is complete. This sequence has a subsequence that converges to some element z in K. The sequence itself must converge to z.

Thus, we can get a full characterization of compactness by adding an additional property to completeness. It can be stated that every compact metric space is 'totally bounded'. In fact, a metric space is compact if and only if it is both complete and totally bounded.

12.8 CONNECTED METRIC SPACE

A connected subset of a metric space is one that can be considered as one whole piece. In $\mathbb{R}$, for instance, we like to think of (0, 1) as connected and [0, I] U [2, 3)as disconnected. [0, 1] U [2, 3) is disconnected in the sense there is a gap between them. That is it cannot be considered as one piece. Whereas when we consider (0, 1), it is one piece.

There ideas are formalised in the following definition.

Definition (Connected space): Let (X, d) be a metric space. We say that X is connected if there do not exist two nonempty and disjoint open sets A and B in X such that $X = A \cup B$. If there exists two disjoint non-empty open sets A and B such that $X = A \cup B$, then X is called disconnected and the pair $\{A, B)$ is called a **disconnection of X**.

Let us see some examples.

Example 12.10 Let us consider $\mathbb{R}$ to be the metric space of all non-zero real numbers with the usual metric $d(x, y) = |x - y|$. If $A = \{x \in X : x > 0\}, A = \{x \in X : x < 0\}$, then the pair $\{A, B\}$ is a disconnection of X.

Example 12.11 Let us consider $\mathbb{Q}$, the metric space of all rational numbers with the usual metric $d(x, y) = |x - y|$. Let r be any irrational number. If $A = \{x \in X : x > r\}$, and $A = \{x \in X : x < r\}$, then the pair $\{A , B)$ is a **disconnection of X**.

12.8.1 Theorems Based on Connectedness of Metric Space

Theorem 1: A subset E of a metric space X is a connected set in X if E is notcontained in the union of two open subsets of X whose intersection with E arenon-empty and disjoint.

Proof : Let $E \in A \cup B$ where A, B are open subsets of X with $A \cap E \neq \varphi, B \cap E \neq \varphi, A \cap B \cap E \neq \varphi$. Then the pair $\{A \cap E, B \cap E)$ becomes a disconnection of E.

Theorem 2: Let f be a continuous mapping of a metric space X onto a metricspace Y. If X is connected, then Y is connected.

Proof: Let X be connected. Let A be a subset of Y which is both open and closed. Continuity of f implies that $f^{-1}(A)$ is both open and closed in X. Connectedness of X implies that either $f^{-1}(A) = \varphi$ or $f^{-1}(A) = X$. Since, $f^{-1}(A) = \varphi$ implies $A = \varphi$ and $f^{-1}(A) = X$ implies $A = Y$. Hence, Y is connected.

Theorem 3 : A metric space X is connected if and only if every continuous function f on X to the discrete metric space $\{1, -1\}$ is a constant function.

Proof : Let Y be the discrete metric space (1, – 1). Let X be connected and $f : X \rightarrow \{+ 1. -1\}$ be continuous. Suppose that f is not a constant. Let $A = f^{-1}\{1)$ and $B = f^{-1}\{-1)$. Then A and B are disjoint non-empty open subsets of X satisfying $X = A \cup B$. This is a contradiction to the fact that X is connected. Therefore f is constant.

To get the reverse implication, let us assume that X is not connected. Let $\{A, B)$ be a disconnection of X. Now define a map $f : X \rightarrow Y$ as

$$f(x) = 1 \text{ if } x \in A$$
$$= -1 \text{ if } x \in B$$

Then f is continuous because the only open subsets of Y are φ, {I}, {–1} and Y. Also f is not a constant.

Another way of stating the above theorem is as follows:

A metric space X is disconnected if and only if there exists a continuous function from X onto a two-point discrete metric space.

Theorem 4 Let X be a metric space. Let A and B be two connected subsets of X such that $A \cap B \neq \varphi$. Then, $A \cup B$ is said to be connected.

12.8.2 Few Examples of Connected Metric Spaces

1. *Closed Interval in* $\mathbb{R}$: The closed interval $[a, b]$ in the real line $\mathbb{R}$ is a connected metric space. Any attempt to split it into two disjoint open sets results in an open set that is not entirely contained within the interval.
2. *Real Line* $\mathbb{R}$: The entire real line $\mathbb{R}$ is a connected metric space. There are no disjoint open sets whose union covers the entire real line.
3. *Open Interval in* $\mathbb{R}$: The open interval (a, b) in the real line $\mathbb{R}$ is a connected metric space. It is not possible to split it into two disjoint open sets whose union forms the entire interval.
4. *Euclidean Spaces* $\mathbb{R}^n$: Euclidean spaces $\mathbb{R}^n$ are connected. This includes spaces such as $\mathbb{R}^2$ and $\mathbb{R}^3$
5. *Circle in* $\mathbb{R}^2$: The circle in the Euclidean plane $\mathbb{R}^2$ is a connected metric space.
6. *Closed Unit Ball in* $\mathbb{R}^n$: The closed unit ball in Euclidean space $\mathbb{R}^n$ is connected.
7. *Path-Connected Spaces*: Spaces that are not only connected but also path-connected, such as the unit circle or any convex set in Euclidean space, where there is a continuous path between any two points.
8. *Graphs of Continuous Functions*: The graph of a continuous function $f: \mathbb{R} \rightarrow \mathbb{R}$ is a connected metric space.

These examples illustrate the different types of connected metric spaces, ranging from intervals in $\mathbb{R}$ to higher-dimensional Euclidean spaces and specific geometric shapes. Connectedness is a fundamental concept in topology, and these examples showcase its applicability in different contexts.

12.8.3 Problems on Connected Metric Space

Problem 1 Prove that the open interval (a, b) in $\mathbb{R}$ is connected.

Solution: Proof: Assume, for contradiction, that (a, b) is not connected. Then, there exist disjoint open sets U and V in $\mathbb{R}$ such that $(a, b) \subset U \cup V$. Without loss of generality, assume that $a \in U$ and $b \in V$. This leads to a contradiction, thus proving that (a, b) is connected.

Problem 2 Let X be a metric space, and let $A \subset X$ be a connected set. Prove that the closure $\bar{A}$ is also connected.

Solution: Proof: Assume, for contradiction, that $\bar{A}$ is not connected. Then, there exist disjoint open sets U and V in X such that $\bar{A} \subset U \cup V$. This again leads to a contradiction, demonstrating that $\bar{A}$ is connected.

Problem 3 Let X be a metric space. Prove that a singleton set $\{x\}$ is connected.

Solution: Proof: A singleton set $\{x\}$ has only one point. Since there are no disjoint nonempty open sets whose union covers $\{x\}$, $\{x\}$ is connected.

PROBLEMS

1. How can you characterize an open and closed set on the real number line?
2. If the intersection of two sets is empty, can they both be open?
3. Which norm is also known as the 'taxicab' norm?
4. What is the role of the Cauchy sequence concept in calculus and analysis?
5. Why is completeness important in the context of Cauchy sequences?
6. In what kind of spaces does the Cauchy completeness property hold?
7. How can you characterize a closed set on the real number line?
8. Give an example of a set that is closed but not compact.
9. Prove that every compact subset of a metric space is closed and bounded. Prove that a closed subset of a compact space is compact.
10. Let (X, d) be a complete metric space, and $Y \subset X$. Prove that (Y, d) is complete if and only if Y is a closed subset of X.
11. What is the relationship between boundedness and compactness?
12. What is the relationship between a connected metric space and the connectedness of its subsets?
13. Give an example of an incomplete metric space. If a metric space is not complete, what can be said about the existence of limits of certain sequences?
14. What role does completeness play in the construction of real numbers?

☐☐☐

CHAPTER

13

VECTOR AND VECTOR SPACE

Chapter Outline

- Vector • Vector addition and Scaler multiplication • Hyperplanes and lines in $\mathbb{R}^n$ • Directions and Magnitudes • Field • Vector Space • Spanning Sets • Subspaces • Linear Dependence and Independence • Basis and Dimensions

13.1 VECTOR

A vector V can be considered as an ordered list of numbers. In other words, a vector is a mathematical entity characterized by both magnitude and direction, often represented as an ordered set of values in a multi-dimensional space.

$$V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

Above vector is an example of a n-dimensional vector, or n-dimensional column vector. The set of all n-tuples of real number is denoted as $\mathbb{R}^n$. As an equation it is written as,

$$\mathbb{R}^n = \left\{ \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \middle| v_1, \dots v_n \in \mathbb{R} \right\}$$

Thus, a particular n-tuple in $\mathbb{R}^n$, say $V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$ denote a point in n-space. The number

v_1 is called as the coordinates, components, entries, or elements of V. Another vector $U = (u_1, u_2, ..., u_m)$ is an example of an m-dimensional row vector.

$$\mathbb{R}^m = \{(u_1, u_2, ..., u_m) \mid u_1, u_2, ..., u_m \in \mathbb{R}\}$$

13.2 VECTOR ADDITION AND SCALER MULTIPLICATION

Consider two vectors U and V in $\mathbb{R}^n$, $U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$ and $V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$, then the sum, written as $U + V$ is the vector in $\mathbb{R}^n$, obtained by adding the corresponding elements from U and V.

That is, $U + V = \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_n + v_n \end{bmatrix}$

The scalar product of the vector V by a real number k is obtained by multiplying each element of V.

By k, That is, $$kV = k\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} kv_1 \\ kv_2 \\ \vdots \\ kv_n \end{bmatrix}$$

Example 13.1

(*a*) Let $A = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$ and $B = \begin{bmatrix} 4 \\ 3 \\ 2 \\ 0 \end{bmatrix}$, then $A + B = \begin{bmatrix} 1+4 \\ 2+3 \\ 3+2 \\ 4+0 \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \\ 5 \\ 4 \end{bmatrix}$

and $3A - B = \begin{bmatrix} 3(1)-4 \\ 3(2)-3 \\ 3(3)-2 \\ 3(4)-0 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ 7 \\ 12 \end{bmatrix}$

(*b*) Let

$$\begin{aligned} A &= (2, 4, -5) \text{ and } B = (1, -6, 9), \text{ then} \\ A + B &= (2 + (1), 4 + (-6), -5 + (9)) = (3, -2, 4), \\ 5A &= (5(2), 5(4), 5(-5)) = (10, 20, -25) \\ -B &= (-1)(1, -6, 9) = (-1, 6, -9), \text{ and} \\ 3A - 5B &= 3(2, 4, -5) + (-5)(1, -6, 9) = (1, 42, -60), \end{aligned}$$

(c) A special vector is the zero vector. All its elements are zero, as : $0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$. The zero vector labels the origin. In this sense, the zero vector is the only one with zero magnitude, only the only one which points in no direction. For any vector, $A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^n$, $A + 0$

$$= \begin{bmatrix} a_1 + 0 \\ a_2 + 0 \\ \vdots \\ a_n + 0 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = A.$$

13.3 HYPERPLANES AND LINES IN $\mathbb{R}^n$

A hyperplane in $\mathbb{R}^n$ is defined as the solution set of a linear equation of the form

$$a_1x_1 + a_2x_2 + \dots a_nx_n = b$$

where, $A = (a_1, a_2, \dots, a_n)$ is a non-zero constant vector in $\mathbb{R}^n$, $x_1, x_2, \dots, x_n \in \mathbb{R}^m$ are variables representing the coordinates in $\mathbb{R}^m$, and b in $\mathbb{R}^m$ is a constant.

Thus, a hyperplane in $\mathbb{R}^2$ is a line and a hyperplane in $\mathbb{R}^3$ is a plane. Given two non-zero vectors. U and V, they will usually determine a plane, unless both vectors are in the same line. The line L in $\mathbb{R}^n$ along the direction of a non-zero vector $U = (u_1, u_2, \dots \dots u_n)$ and passing through a point $P(b_1, b_2, \dots, b_n)$ can be written as:

$$L = (P + tU \mid t \in \mathbb{R}).$$

The plane determined by two vectors U and V in $\mathbb{R}^n$ can be written as:

$$\{P + tU + sV \mid t, s \in \mathbb{R}\}$$

Example 13.2

(a) In a two-dimensional scenario, $\left\{ \begin{bmatrix} 1 \\ 2 \\ 5 \\ 4 \end{bmatrix} + t \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \middle| t \in \mathbb{R} \right\}$ describe a line in $\mathbb{R}^4$ parallel to the x-axis.

(*b*) In a three-dimensional scenario, $\left\{\begin{bmatrix}2\\1\\4\\1\\5\\8\end{bmatrix}+s\begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}+t\begin{bmatrix}0\\1\\0\\0\\0\\0\end{bmatrix}\middle| s,t\in\mathbb{R}\right\}$ describes a plane in $\mathbb{R}^6$ parallel to *x*-*y* plane.

(*c*) (Specifying a plane with one linear algebraic equation); The solution set to $x_1 + x_2 + x_3 + x_4 + x_5 = 1$ can be represented equivalently as:

$$\left\{\begin{bmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{bmatrix}=\begin{bmatrix}1-x_2-x_3-x_4-x_5\\x_2\\x_3\\x_4\\x_5\end{bmatrix}\right\}$$

, which further can be written as hyperplane,

$$\left\{\begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}+s_2\begin{bmatrix}-1\\1\\0\\0\\0\end{bmatrix}+s_3\begin{bmatrix}-1\\0\\1\\0\\0\end{bmatrix}+s_4\begin{bmatrix}-1\\0\\0\\1\\0\end{bmatrix}+s_5\begin{bmatrix}-1\\0\\0\\0\\1\end{bmatrix}\middle| s_2,s_3,s_4,s_5\in\mathbb{R}\right\}$$

Thus, the solution set to $x_1 + x_2 + x_3 + x_4 + x_5 = 1$ is 4-dimensional hyperplane in $\mathbb{R}^5$.

Example 13.3

(*a*) Let *H* be the plane in $\mathbb{R}^3$ corresponding to the linear equation $2x_1 - 5x_2 + 7x_3 = 4$.

Notice that the point $P = (1, 1, 1)$ and $Q = (5, 4, 2)$ are the solutions of the equation.

Thus, *P* and *Q* and the directed line segment $V = \overline{PQ} = Q - P = (4, 3, 1)$ lie on the plane *H*. The vector $U = (2, -5, 7)$ is normal to *H*, and as expected,
$U \cdot V = (2, -5, 7) \cdot (4, 3, 1) = 8 - 15 + 7 = 0$, Therefore, *U* is orthogonal to *V*.

(*b*) Find an equation of the hyperplane *H* in $\mathbb{R}^4$ that passes through the point $P = (1, 3, -4, 2)$ and is normal to the vector $U = (4, -2, 5, 6)$. The coefficients of the unknown the equation of *H* are the components of the normal vector *U*; hence the equation of *H* must be of the form $4x_1 - 2x_2 + 5x_3 + 6x_4 = k$.

Substituting P into this equation, we obtain $4(1)\ \ 2(3) + 5(-4) + 6(2) = k$ or $k = -10$.

Thus, $4x_1 - 2x_2 + 5x_3 + 6x_4 = -10$, is the equation of hyperplane H.

13.4 DIRECTIONS AND MAGNITUDES

Definition: The dot product, also known as inner product or scaler product of vectors in $\mathbb{R}^n$,

$$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \text{ is defined as}$$

$$U \cdot V = u_1v_1 + u_2v_2 + \dots + u_nv_n$$

That is the dot product of vectors U and V is obtained by multiplying corresponding elements and adding the resulting products.

The vectors U and V are said to be perpendicular (or orthogonal) if their dot product is zero, that is $U \cdot V = 0$. The product is also denoted as $\langle U, V\rangle$.

Example 13.4

(*a*) Let $$U = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}, V = \begin{bmatrix} 2 \\ -3 \\ 5 \end{bmatrix}, \quad \text{and} \quad W = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}, \text{ then}$$

$$U \cdot V = 1(2) + 2(-3) + (-2)(5) = 2 - 6 - 10 = -14$$

$$U \cdot V = W = 1(2) + 2(3) + (-2)(4) = 2 + 6 - 8 = 0$$

$$V \cdot W = 2(2) + (-3)(3) + 5(4) = 4 - 9 + 20 = 15$$

Thus, U and V are orthogonal.

(*b*) Let $U = (1, 2, 3, 5)$ and $V = (2, 3, k, 5)$. Then find k so that U and V perpendicular.

We have, $U \cdot V = 1(2) + 2(3) + 3k + 5(5) = 2 + 6 + 3k + 25 = 33 + 3k$

U and V are perpendicular then $U \cdot V = 0$.

Therefore, $33 + 3k = 0$, or $k = -11$.

Some Properties of Dot Product: Consider the non-zero vectors U, V and W in $\mathbb{R}^n$. Then, some important properties of dot products are mentioned as below:

(i) Symmetry: $U \cdot V = V \cdot U$

(ii) Distributive: $U \cdot (V + W) = U \cdot V + U \cdot W$

(iii) Bilinear (linear in both U and V): $U \cdot (cU + dW) = cU \cdot V + dU \cdot W)$ and $(cU + dW) \cdot V = cU \cdot V + dW \cdot V$, where c, d are scaler in $\mathbb{R}$.

(iv) Positive Definite: $U \cdot U > 0$.

Euclidean length (Norm or Magnitude) of a Vector:

The Euclidean norm of a vector $V \in \mathbb{R}^n$, is denoted by $\|V\|$, and defined to be the non-negative square root of $\langle V, V \rangle$. Consider a vector $V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$ in $\mathbb{R}^n$, then

$$\|V\| = \sqrt{\langle V,V \rangle} = \sqrt{(v_1)^2 + (v_2)^2 + \ldots (v_n)^2} = \sqrt{\sum_{i=1}^{n} (v_i)^2}$$

Thus, $\|V\| \geq 0$, and $\|V\| = 0$ if and only if $V = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$. A vector is called unit vector (or unit norm vector) if $\|V\| = 1$, or equivalently if $\langle V, V \rangle = 1$.

For any non-zero vector $V \in \mathbb{R}^n$, the vector $\hat{V} = \dfrac{V}{\|V\|}$ is the unit norm vector in the direction of V. The process of obtaining $\hat{V}$ from V is called the normalization of vector V.

Distance, Angles and Projection: Consider two non-zero vectors U and V in $\mathbb{R}^n$,

$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$ and $V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$, then the ***distance*** between U and V is denoted and defined as:

$$d(U, V) = \|U - V\| = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \ldots (u_n - v_n)^2}$$

The angle between U and V is determined by the formula,

$$U \cdot V = \|U\| \, \|V\| \cos \theta \text{ or } \cos \theta = \frac{U \cdot V}{\|U\| \; \|V\|}$$

The projection of a vector U onto a non-zero vector V is denoted and determined by,

$$\text{Proj}(U, V) = \frac{U \cdot V}{\|V\|^2} V = \frac{U \cdot V}{V \cdot V} V$$

Theorem 1: (Cauchy-Schwarz Inequality): For any non-zero vectors U and V in $\mathbb{R}^n$,

$$U \cdot V \leq \|U\| \, \|V\|$$

Proof: Let $\alpha \in \mathbb{R}^n$ any real number and consider the following positive quadratic polynomial in α.

$$\begin{aligned} 0 \leq (U + \alpha V) \cdot (U + \alpha V) &= U \cdot U + 2\alpha U \cdot V + \alpha^2 V \cdot V \\ &= \|U\|^2 + 2(U \cdot V)\alpha + \|V\|^2 \alpha^2 \end{aligned}$$

Let, $a = \|V\|^2$, $b = 2(U \cdot V)$, and $c = \|U\|^2$. Then for every value of α

$$a\alpha^2 + b\alpha + c \geq 0.$$

This means that the determinant $D = b^2 - 4ac \leq 0$.

Equivalently, $b^2 \leq 4ac$. Thus, $4(U \cdot V)^2 \leq 4\|U\|^2 \, \|V\|^2$ or $U \cdot V \leq \|U\| \, \|V\|$

Theorem 2: Triangle inequality: for any vectors U and V in $\mathbb{R}^n$,

$$\|U + V\| \leq \|U\| + \|V\|.$$

Proof:

$$\begin{aligned} \|U + V\|^2 &= (U + V) \cdot (U + V) \\ &= U \cdot U + 2U \cdot V + V \cdot V \\ &= \|U\|^2 + \|V\|^2 + 2\|U\| \, \|V\| \cos\theta \\ &= (\|U\| + \|V\|)^2 + 2\|U\| \, \|V\| (\cos\theta - 1) \\ &\leq (\|U\| + \|V\|)^2 \end{aligned}$$

Thus, $\|U + V\| \leq \|U\| + \|V\|$

The triangle inequality is also self-evident when examining a sketch of u, v and $u + v$ as below:

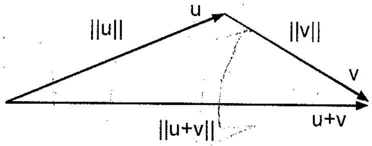


Fig. 13.1

Example 13.5.

Consider two vectors in $\mathbb{R}^3$, $U = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$ and $V = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$. Then obtain the distance and angle between U and V.

Distance between U and V is obtained as

$$d(U, V) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \dots (u_n - v_n)^2}$$

$$= \sqrt{(1-2)^2 + (-2-4)^2 + (3-6)^2}$$

$$= \sqrt{1+36+9} = \sqrt{46}$$

To find the angle between U and V, we first obtain

$$U \cdot V = (1)(2) + (-2)(4) + (3)(6) = 12,$$

$$||U||^2 = (1)^2 + (-2)^2 + (3)^2 = 14,$$

$$||V||^2 = (2)^2 + (4)^2 + (6)^2 = 56,$$

Then, $$\cos\theta = \frac{U \cdot V}{||U|| \ ||V||} = \frac{12}{\sqrt{4}\sqrt{56}} = \frac{3}{7} \quad \text{or } \cos^{-1}\left(\frac{3}{7}\right)$$

Also, $$\text{proj}(U, V) = \frac{U \cdot V}{||V||^2} V = \frac{12}{56}\begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} = \frac{3}{7}\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

13.5 FIELD

In the context of vector spaces, a field F is a mathematical structure that provides the scalars used for scalar a multiplication in the vector space. A field is a set equipped with two operations, addition and multiplication, such that the set satisfies certain algebraic properties, the following properties hold for all elements a, b, c in the field.

dkl

- (i) **Closure under addition and multiplication:** $a + b$ and $a \cdot b$ are in field F.
- (ii) **Associativity of addition and multiplication:** $a + (b + c) = (a + b) + c$. and $a \cdot (b \cdot c) + (a \cdot b) \cdot c$.
- (iii) **Commutativity of addition:** $a + b = b + a$.
- (iv) **Existence of additive identity:** There exists an element 0 in F such that, $a + 0 = a$ for all a in F.
- (v) **Existence of additive inverses:** For every a in F, there exists an element $-a$ in such that $a + (-a) = 0$.
- (vi) **Non-zero element property:** $a \,.\, b = 0$ implies that either $a = 0$ or $b = 0$.
- (vii) **Existence of multiplicative identify:** There exists an element 1 in F such that $a \cdot 1 = a$ for all a in F.

13.6 VECTOR SPACE

Formally, a vector space over a field (usually the real numbers or complex numbers) is a set V equipped with two operations: vector addition and scalar multiplication, such that for any vectors $u, v \in V$ and any scalars c, d from the field F, the following properties hold:

(i) **Closure under addition:** $u + v \in V$.

(ii) ***Associativity of addition:*** $u + (v + w) = (u + v) + w$.

(iii) **Commutativity of addition:** $u + v = v + u$

(iv) **Existence of zero vector:** There exists a vector $0 \in V$ such that $u + 0 = u$ for all $u \in V$.

(v) **Existence of additive inverse:** For every $u \in V$, there exists a vector $-u \in V$ such that $u + (-u) = 0$.

(vi) **Closure under scalar multiplication:** $c \cdot u \in v$.

(vii) ***Compatibility of scalar multiplication with field multiplication:*** $c(d \cdot u) = (cd \cdot u)$

(viii) **Identity element for scalar multiplication:** $1 \cdot u = u$ for all $u \in V$.

(ix) **Distributivity of scalar multiplication with respect to vector addition:**

$$c \cdot (u + v) = c \cdot u + c \cdot v$$

(x) **Distributivity of scalar multiplication with respect to field addition:**

$$(c + d) \cdot u = c \cdot u + d \cdot u$$

13.7 SPANNING SETS

Let V be a vector space over F. Then vectors $u_1, u_2, \ldots, u_n$ are said to span V or form a spanning set of V if every $v \in V$, a linear combination of the vectors $v_1, v_2, \ldots, v_n$. That means if there exist scaler $a_1, a_2, \ldots a_n \in F$, then, $v = a_1u_1 + a_2u_2 + \ldots + a_nu_n$ belongs to V as well. The following remarks can be concluded from the above definition.

Remark 1: Suppose $u_1, u_2, \ldots, u_n$ span V. Then, for any vector w, the set $w, u_1, u_2, \ldots, u_n$ also spans V.

Remark 2: Suppose $u_1, u_2, \ldots, u_n$ span V and suppose u_k is a linear combination of some of the other u's. Then the u's without u_k also span V.

Remark 3: Suppose $u_1, u_2, \ldots, u_n$ span V and suppose one of the u's is the zero vector. Then the u's without the zero vector also spans V.

Example 13.6 Let $V = \mathbb{R}^2$(the vector space of all 2-dimenstional real vectors) and consider the set $S = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$. The question is whether S spans V. Any vector in $\mathbb{R}^2$ can be written as a linear combination of the vectors in S. For example:

The vector $v_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ can be expressed as $v_1 = 2\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

Since we can represent any vector in $\mathbb{R}^2$ using linear combinations of the vectors in S, we can say that S span $\mathbb{R}^2$.

Example 13.7 Now let's consider $V = \mathbb{R}^3$ and the set $S = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\}$

Again, we want to know if S spans V. For a vector in $\mathbb{R}^3$, say (a, b, c), we need to check if there exist scalars x, y and z such that:

$$x\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + y\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + z\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \qquad \text{or, } x + y + z = a;\ x + y = b;\ x = c;$$

In this case, we can see that the vector $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ can be written as

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = c\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + (b-c)\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + (a-b)\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

So, it is concluded that S spans $\mathbb{R}^3$. These examples illustrate the concept of spanning sets, where a set of vectors spans a vector space if every vector in that space can be expressed a linear combination of the vectors in the set.

13.8 SUBSPACES

In the realm of subspace geometry, envision the conventional three-dimensional space, $\mathbb{R}^3$, and select any plane passing through the origin. This chosen plane constitutes a distinct vector space. When we scale a vector within this plane by a factor like 2, – 3 or any scalar, the result remains within the same plane. Similarly, the sum of two vectors within the plane preserves the plane. This specific plane, passing through the origin $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, exemplifies a foundational concept in linear algebra-it serves as a subspace within the original space $\mathbb{R}^3$.

Definition: Subspace S of a vector V is a non empty subset that satisfies the requirements for a vector space that Linear combinations stay in the subspace.

(i) The sum of any two vectors, x and y, also resides in the subspace: $(x + y) \in S$.

(ii) The product of any vector x in the subspace by any scalar c remains within the subspace: $cx \in S$.

Example 13.8 Consider all vectors in $\mathbb{R}^2$, whose components are positive or zero. This subset is the first quadrant of the x-y plane; the coordinates satisfy $x \geq 0$. and $y \geq 0$. It is not a subspace, even though it contains zero and addition does leave us within the subset. Rule (ii) is a violated, since if the scalar $c = -1$ and the vector $x = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, then the multiple $cx = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$ is in the third quadrant instead of the first. If we include the third quadrant along with the first, scalar multiplication is all right. Every multiple cx will stay in this subset. However, rule (i) is now violated, since $\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} -2 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, which is not in either quadrant. The smallest subspace containing the first quadrant is the whole space $\mathbb{R}^2$.

Example 13.9 Start from the vector space of 3 × 3 matrices say $\mathbb{R}^{3 \times 3}$. One possible subspace is the set of *lower triangular matrices*. Another is the set of *symmetric matrices*. Therefore, $A = B$ and cA are lower triangular if A and B are lower triangular, and they are symmetric if A and B are symmetric. Of course, the zero matrix is in both subspaces.

Let's see two fundamental examples: the **column space** and the **null-space** of a matrix A.

Column space: The column space encompasses every possible linear combination that can be formed using the columns of matrix A.

Explanation: The column space of a matrix refers to the subspace formed by all the possible linear combinations of its columns. Consider a matrix A with columns $[v_1, v_2, v_3]$. The column space would include every vector that can be expressed as a linear combination of these columns, such as $c_1v_1 + c_2v_2 + c_3v_3$ where c_1, c_2 and c_3 are scalars. In essence, it represents the span of the column vectors of the matrix.

Example 13.10 If $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$, then the column space would include all vectors of the form. $c_1\begin{bmatrix}1\\3\\5\end{bmatrix}+c_2\begin{bmatrix}2\\4\\6\end{bmatrix}$ where c_1 and c_2 can be any scalar. The column space is the subspace

spanned by the columns $\begin{bmatrix}1\\3\\5\end{bmatrix}$ and $\begin{bmatrix}2\\4\\6\end{bmatrix}$ in this example.

Note: The system of linear equation, say $Ax = b$ have a solution if and only if the vector b can be represented as a linear combination of the columns of matrix A. In other words, b lies within the column space of A.

Nullspace: The null space of a matrix, denoted as $N(A)$, comprises all vectors x that satisfy the equation $Ax = 0$. Therefore, the null space of a matrix A represents the set of all solutions (vectors x) that, when multiplied by A, result in the zero vector ($Ax = 0$).

In other words, it consists of vectors that get 'mapped' to the zero vector under the linear transformation represented by the matrix A.

This null space is a subspace of the vector space $\mathbb{R}^n$, analogous to how the column space is a subspace of $\mathbb{R}^m$.

Example 13.11 Let's consider a specific example of a matrix A and find its null space in $\mathbb{R}^3$:

$$A = \begin{bmatrix}1 & 2 & -1\\0 & 1 & 1\\2 & 0 & 3\end{bmatrix}$$

Now, we are looking for vectors $X = \begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}$ such that $Ax = b$.

Setting up the system of equations: $\begin{bmatrix}1 & 2 & -1\\0 & 1 & 1\\2 & 0 & 3\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}$

Solving this system, you can find the solutions for x_1, x_2, x_3. The null space ($N(A)$) consists of all vectors of the form $\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}$ that satisfy these equations. The set of solutions forms the basis for the null space of matrix A in $\mathbb{R}^3$.

13.9 LINEAR DEPENDENCE AND INDEPENDENCE

Linear dependence and independence are important concepts that describe the relationship between vectors.

1. Linearly Dependent Vectors: A set of vectors is said to be linearly dependent if there exists a nontrivial linear combination of these vectors that equals the zero vector. In simpler terms, vectors, $v_1, v_2, ..., v_n$ are linearly dependent if there exist scalars $c_1, c_2, ..., c_n$ not all zero such that $c_1v_1 + c_2v_2 + ... c_nv_n = 0$.

Example 13.12 Consider two vectors in $\begin{bmatrix}1\\2\end{bmatrix}$ and $v_2 = \begin{bmatrix}2\\4\end{bmatrix}$

These vectors are linearly dependent because $2v_1 = v_2$.

You can verify that, $2\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}2\\4\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$.

2. Linearly Independent Vectors: A set of vectors is said to be linearly independent if the only linear combination that equals the zero vector is the trivial one, where all the coefficients are zero.

In other words, vectors $v_1, v_2, ..., v_n$ are linearly independent if $c_1v_1 + c_2v_2 + ... c_nv_n = 0$ implies that $c_1 = c_2 = c_n = 0$.

Example 13.13 Consider two vectors in $\mathbb{R}^2$: $v_1 = \begin{bmatrix}1\\0\end{bmatrix}$ and $v_2 = \begin{bmatrix}0\\1\end{bmatrix}$.

These vectors are linearly independent because the only way to get the zero vector as a linear combination is by setting $c_1 = c_2 = 0$, since $c_1\begin{bmatrix}1\\0\end{bmatrix} + c_2 = \begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$

In general, if you have a set of vectors $v_1, v_2, ..., v_n$ and you can express one of them as a linear combination of the others, the set is linearly dependent. If no vector in the set can be written as a linear combination of the others, the set is linearly independent. Linear independence is a desirable property because it means that none of the vectors in the set is redundant; each contributes somethings unique to the span of the set.

13.10 BASIS AND DIMENSIONS

A basis is a fundamental concept in the context of vector spaces. A basis is a set of vectors that spans the entire vector space and is linearly independent. The idea is that any vector in the vector space can be uniquely expressed as a linear combination of the vectors in the basis. The number of vectors in the basis is known as the dimension of the vector space. Some key properties of a basis is mentioned below:

1. **Spanning Property:** A basis must span the entire vector space, meaning that any vector in the space can be represented as a linear combination of the basis vectors.
2. **Linear Independence:** The vectors in a basis must be linearly independent, ensuring that no vector in the basis can be written as a combination of the others.

Example 13.14: Consider the vector space $\mathbb{R}^3$, and let's define a basis for this space. A common basis for $\mathbb{R}^3$ is the standard basis, which consists of three vectors:

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \text{and } e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

These vectors form a basis for $\mathbb{R}^3$ because:

Spanning Property: Any vector $V = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ in $\mathbb{R}^3$ can be expressed as a linear combination of the standard basis vectors: $V = a \,.\, e_1 + b \,.\, e_2 + c \,.\, e_3$

Linear Independence: The coefficients in the linear combination above are unique. If $a \,.\, e_1 + b \,.\, e_2 + c \,.\, e_3 = 0$, then $a = b = c = 0$. This ensures linear independence.

Now, let's take an example vector $W = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$. Then we can express W as a linear combination of the standard basis vectors, as : $W = 2 \,.\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 3 \,.\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 1 \,.\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

This demonstrates the spanning property of the basis, showing that any vector in $\mathbb{R}^3$ can be represented using the basis vectors.

Thus, we can say that a basis is a set of vectors that not only spans the vector space but also ensures that the vectors are linearly independent, providing a unique way to represent any vector in the space. The standard basis for $\mathbb{R}^n$ is a common example that satisfies these properties.

PROBLEMS

1. Let $U_1 = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}, U_2 = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix}$, and $U_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$

(a) $3U_1 - 2U_2$ (b) $2U_2 + U_2 - 7U_3$

2. Write the vector $V = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ as a linear combination of the vectors $U_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, U_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$, and $U_3 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$.

3. Find $U \cdot V$ where :

(a) $U = (2, 3, 4, 5)$ and $V = (3, -2, 5, 4)$

(b) $U = (2, -3, 4-5, 1)$ and $V = (3, -1, -5, 4, , 2)$

4. Consider $U = (3, 2, -2, 1)$ and $V = (3, k, 5, 4)$. The find the value of k such that U and V are orthogonal.

5. Find the unit vector in the direction of vector V where:

(a) $V = (3, -2, 2, 4)$

(b) $V = (2, -3, 4, 5, 1)$

6. Consider $U = (1, -2, 4)$ and $V = (-3, 2, 5)$. Then find:

(a) $\text{Cos}\ \theta$ where θ is the angle between U and V.

(b) proj (U, V), the projection of U onto V.

(c) $d(U, V)$, the distance between U and V.

7. Find an equation of the hyperplane H that passes through points $P = (3, -4, 2)$ and is normal to vector $V = (-3, 2, 5)$.

8. Check whether the vectors: $U_1 = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}, U_2 = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix}$, and $U_3 = \begin{bmatrix} 3 \\ -3 \\ -1 \end{bmatrix}$ are linearly independent.

9. Determine whether the vectors: $U_1 = \begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix}, U_2 = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix}$, and $U_3 = \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix}$ spans $\mathbb{R}^3$.

□□□

CHAPTER

14

EIGEN VALUES AND EIGEN VECTORS

Chapter Outline

• Introduction • Cayley-Hamilton theorem, Characteristic Polynomial • Eigen Values and Eigen Vectors • Some other Properties of Eigen values • Diagonalization of a Matrix

14.1 INTRODUCTION

Eigen values and eigen vectors provide a way to understand the inherent properties of linear transformations represented by matrices. In linear algebra statistical modelling, eigen values and eigen vectors are utilized to uncover the principal components of a dataset, enabling dimensionality reduction, and capturing the most significant variability in the data.

Consider an example of multiplying the non-zero vectors $\begin{bmatrix} 2 \\ 5 \end{bmatrix}$ and $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$ by a given square matris $\begin{bmatrix} 6 & 3 \\ 4 & 7 \end{bmatrix}$ such as

Case 1. $\begin{bmatrix} 6 & 3 \\ 4 & 7 \end{bmatrix}\begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 27 \\ 43 \end{bmatrix}$

Case 2. $\begin{bmatrix} 6 & 3 \\ 4 & 7 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 30 \\ 40 \end{bmatrix}$

We aim to examine the impact of multiplying the given matrix on vectors. In case 1. It is observed that the result is a completely different vector with altered direction and length, which is typically unremarkable for our current discussion. However, in the case 2, something noteworthy unfolds. The multiplication yields a vector, signifying that the new vector maintains the same direction as the original one. The scaling factor denoted as λ is 10. This chapter will delve into the systematic exploration of such scaling factors λ and non-zero vectors X for a given square matrix A by considering the equation of the form $AX = \lambda X$.

14.2 CAYLEY-HAMILTON THEOREM, CHARACTERISTIC POLYNOMIAL

The Cayley-Hamilton theorem is a fundamental result in linear algebra that establishes a relationship between a square matrix and its characteristic polynomial. The theorem is named after the mathematicians Arthur Cayley and William Rowan Hamilton.

Consider a square matrix A of order $n \times n$. Then the characteristic polynomial of A is denoted by $p(\lambda) = \det(A - \lambda I)$, where I is the identity matrix and λ is a scaler. And the characteristic polynomial in λ whose roots are the eigenvalues of the matrix A.

A matrix polynomial is an expression of the form $P(\lambda) = c_0 I + c_1\lambda + c_2\lambda^2 + \ldots + c_k\lambda^k$, where $c_0, c_1, \ldots, c_k$ are constants. The degree of the matrix polynomial is the highest power of λ that appears in the expression.

Cayley-Hamilton Theorem: The Cayley-Hamilton states that every square matrix satisfies its own characteristic equation. In other words, if $p(\lambda) = \det(A - \lambda I)$, is the characteristic polynomial of the matrix A, then substituting A for λ in the polynomial yields the zero matrix. That is, $p(A) = \det(A - AI) = 0$.

For a matrix polynomial $P(\lambda) = c_0 I + c_1\lambda + c_2\lambda^2 + \ldots + c_k\lambda^k$, the Cayley-Hamilton theorem implies:

$$P(A) = c_0 I + c_1 A + c_2 A^2 + \ldots + c_k A^k = 0$$

Example 14.1 Consider a 2×2 matrix A given by: $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

Then the characteristic polynomial $p(\lambda)$ is found by evaluating $\det(A - \lambda I)$:

$$p(\lambda) = \det\left(\begin{bmatrix} 1-\lambda & 2 \\ 3 & 4-\lambda \end{bmatrix}\right) = (1-\lambda)(4-\lambda) - 2 \,.\, 3 = \lambda^2 = 5\lambda + 4$$

Now, according to the Cayley-Hamilton theorem for matrix polynomials:

$$P(A) = A^2 - 5A + 4I = 0$$

Substitute, $P(A)$ is the zero matrix, confirming the Cayley-Hamilton theorem for this example.

Example 14.2

(*a*) Let us the find the characteristic polynomial of a 2×2 matrix, $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, The characteristic polynomial of A is given as: $p(\lambda) = \det(A - \lambda I)$

$$\begin{aligned} p(\lambda) &= \det\left(\begin{bmatrix} a_{11}-\lambda & a_{12} \\ a_{21} & a_{22}-\lambda \end{bmatrix}\right) \\ &= (a_{11}-\lambda)(a_{22}-\lambda) - a_{12}.a_{21} \\ &= \lambda^2 - (a_{11}+a_{22})\lambda + (a_{11}\cdot a_{22} - a_{12}\cdot a_{21}) \\ p(\lambda) &= \lambda^2 - tr(A)\lambda + \det(A) \end{aligned}$$

where, $tr(A)$ denotes the trace of A, that is the sum of the diagonal elements of A.

(b) Now, consider a 3×3 matrix, $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$,

With similar approach (as in part (a)), the characteristic polynomial of A is given as:

$$p(\lambda) = \lambda^3 - tr(A)\,\lambda^2 + (A_{11} + A_{22} + A_{33})\lambda - \det(A),$$

Where, A_{11}, A_{22}, A_{33} denotes the cofactors of a_{11}, a_{22} and a_{33} respectively.

14.3 EIGEN VALUES AND EIGEN VECTORS

Definition: Let A be any square matrix. Then a scalar λ is called an eigenvalue of A. If there exists a non-zero (column) vector X such that $AX = \lambda X$ holds. Any vector X satisfying this relation is called an eigen vector of A associated with the eigenvalue λ.

Note that each scalar multiple of an eigen vector X associated with the eigen value λ is also such an eigen vector, as:

$$A(kX) = k(AX) = k(\lambda X) = \lambda(kX).$$

Determining Eigen values and Eigen vectors:

Example 14.3 Consider a 2×2 matrix $A = \begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$ then determine the Eigen values and the eigen vectors of matrix A.

(*a*) To find the Eigen values:

1. Solve the characteristic equation, $\det(A - \lambda I) = 0$.

$$\det\left(\begin{bmatrix} 3-\lambda & 1 \\ 2 & 2-\lambda \end{bmatrix}\right) = (3-\lambda)(2-\lambda) - 2 \times 1 = 0$$

2. Solve for λ, which gives the eigen values.

$$(3-\lambda)(2-\lambda) - 2 = 0$$

$$\lambda^2 - 5\lambda + 4 = 0$$

$$(\lambda - 1)(\lambda - 4) = 0$$

So, the eigen values are $\lambda_1, = 1$ and $\lambda_2 = 4$.

(*b*) To find the eigenvectors:

For each eigenvalue λ, solve the system $(A - \lambda I)X = 0$.

1. For $\lambda = 1$:

$$(A - \lambda I)X = \begin{bmatrix} 3-1 & 1 \\ 2 & 2-1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

One solution is $X_1 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$

2. For $\lambda = 4$:

$$(A - 4I)X = \begin{bmatrix} 3-4 & 1 \\ 2 & 2-4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

One solution is $X_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Therefore, the eigen values are $\lambda_1 = 1$, and $\lambda_2 = 4$, and the corresponding eigenvectors are $X_1 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$ and $X_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Theorem 1: Eigenvalues: The eigenvalues of a square matrix A are the roots of the characteristic equation of A. The, an $n \times n$ matrix has at least one eigenvalue and at most n numerically different eigenvalues.

Theorem 2: Eigenvectors, Eigen space: If W and X are eigenvectors of a matrix A corresponding to ***the same*** eigenvalue λ so are $W + X$ (where, $X \neq - W$) and kX for any $k \neq 0$. Then the eigenvectors corresponding to one and the same eigenvalue of A, together with 0, form a vector space, called the **eigenspace** of A corresponding to that eigenvalue λ.

Example 14.4 (Multiple Eigenvalues:)

Let us find the eigenvalues and eigenvectors of a matrix, $A = \begin{bmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{bmatrix}$

The characteristic equation is given as: det $(A - \lambda I) = 0$.

Or, $\det\left(\begin{bmatrix} -2-\lambda & 2 & -3 \\ 2 & 1-\lambda & -6 \\ -1 & -2 & -\lambda \end{bmatrix}\right) = 0$

Or, $\lambda^3 + \lambda^2 - 21\lambda - 45 = 0$

The roots of above equation gives the eigenvalues of A.

Thus, we get the eigenvalues: $\lambda_1 = 5, \lambda_2 = \lambda_3 = -3$.

The eigenvector X associated with eigenvalue λ is given by $(A - \lambda I)X = 0$.

Where $\lambda = 5$, then $\begin{bmatrix} -2-5 & 2 & -3 \\ 2 & 1-5 & -6 \\ -1 & -2 & -5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

which is row-reduces to, $\begin{bmatrix} -7 & 2 & -3 \\ 0 & -\frac{24}{7} & -\frac{48}{7} \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

Thus, $-7x_1 + 2x_2 - 3x_3 = 0$ and $-\frac{24}{7}x_2 - \frac{48}{7}x_3 = 0$

Let us choose $x_3 = -1$, then we get $x_2 = 2$ and $x_1 = 1$.

Thus, the eigenvector of A corresponding to $\lambda = 5$ is $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}$

Similarly, when $\lambda = -3$, then $\begin{bmatrix} -2+3 & 2 & -3 \\ 2 & 1+3 & -6 \\ -1 & -2 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

which row-reduce to $\begin{bmatrix} 1 & 2 & -3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

thus, $x_1 + 2x_2 - 3x_3 = 0$ or $x_1 = -2x_2 + 3x_3$

Let us choose $x_2 = 1$ and $x_3 = 0$ then we get $x_1 = -2$, that is $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}$

And choosing $x_2 = 0$ and $x_3 = 1$, we get $x_1 = 3$, that is $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}$

Thus, we get two linearly independent eigen vectors corresponding to $\lambda = -3$.

Note that, the order of an eigenvalue λ as a root of the characteristic polynomial is called the **algebraic multiplicity** of λ, and denoted as M_λ. And the number of linearly independent eigenvectors corresponding to is called as the **geometric multiplicity** of λ, denoted as m_λ. This m_λ is the dimension of the eigenspace corresponding to this λ.

Definitions: **(*a*)** A real square matrix A is called symmetric if $A^T = A$.

(*b*) A real square matrix A is called skew-symmetric if $A^T = -A$.

(*c*) A real square matrix A is called orthogonal if $A^T = A^{-1}$.

Theorem 4: (Eigenvalues of the Transpose)

The transpose A^T of a square matrix A has the same eigenvalues as A.

Theorem 5: (Eigenvalues of Symmetric and Skew-Symmetric Matrices)

(*a*) The eigenvalues of a symmetric matrix are real.

(*b*) The eigenvalues of a skew-symmetric matrix are pure imaginary or zero.

Theorem 6: (Invariance of Inner Product)

Consider two vectors $X, Y \in \mathbb{R}^n$ with inner product, $\langle X, Y\rangle = X^TY = [x_1, x_2, \ldots, x_n]\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$,

and a $n \times n$ orthogonal matrix A. Then the orthogonal transformation $U = AX$ and $AV = AY$ preserves the value of the **inner product** of vectors X, Y. That is $\langle U, V\rangle = \langle X, Y\rangle$. Moreover, the transformation also preserves the **length** or **norm** of any vector, that is $\|U\| = \|AX\| = \|X\|$.

Example 14.5 (Real Matrices with Complex Eigenvalues and Eigenvectors)

We know that a real polynomial may have complex roots (which then occur in conjugate pairs), therefore a real matrix may have complex eigenvalues and eigenvectors. For example, Consider a matrix $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, then characteristic equation of this skew-symmetric matrix is $\det(A - \lambda I) = \begin{bmatrix} -\lambda & 1 \\ -1 & -\lambda \end{bmatrix} = \lambda^2 + 1 = 0$. Thus, the eigenvalues are $\lambda_1 = -i$ and $\lambda_2 = i$.

Then the eigenvector $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ corresponding to λ_1 is $\begin{bmatrix} i & 1 \\ -1 & i \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, after solving we get, $ix_1 + x_2 = 0$, let $x_1 = 1$, then $x_2 = i$. Therefore, the eigenvector corresponding to λ_1 is $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ i \end{bmatrix}$. Similarly, the eigenvector corresponding to λ_2 is $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -i \end{bmatrix}$.

Theorem 7: (Any real square matrix $\boldsymbol{A}$ may be written as the sum of a symmetric matrix E and a skew-symmetric matrix O, where E and O are given as: $E = \frac{1}{2}(A + A^T)$ and $O = \frac{1}{2}(A - A^T)$.

Example 14.6: Consider a real square matrix $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$, then $A^T = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$

Thus, $E = \frac{1}{2}(A + A^T) \begin{bmatrix} 1 & 3 & 5 \\ 3 & 5 & 7 \\ 5 & 7 & 9 \end{bmatrix}$, and $O = \frac{1}{2}(A - A^T) = \begin{bmatrix} 0 & -1 & -2 \\ 1 & 0 & -1 \\ 2 & 1 & 0 \end{bmatrix}$.

We observe that, $A = E + O$, where E is a symmetric matrix and O is a skew-symmetric matrix.

Theorem 8: (Eigenvalues of Symmetric and Skew-Symmetric Matrices)

(*a*) The eigenvalues of a symmetric matrix are real.

(*b*) The eigenvalues of a skew-symmetric matrix are pure imaginary or zero.

Theorem 9: (Orthonormality of Column and Row Vectors)

A real square matrix $A \in \mathbb{R}^{n \times n}$ is orthogonal if and only if its column vectors $A_1, A_2, \ldots A_n \in \mathbb{R}^n$(and its row vectors) form an **orthonormal system**, that is,

$$(A_j, A_k) = A_j^T A_k = \begin{cases} 0 & if \quad j \neq k \\ 1 & if \quad j = k \end{cases}$$

Theorem 10: (Determinant of an Orthogoanl Matrix)

The determinant of an orthogonal matrix has the value -1 or $+1$.

Theorem 11: (Eigenvalues of an Orthogonal Matrix)

The eigenvalues of an orthogonal matrix A are either real or complex conjugates in pairs, and have absolute value 1.

Theorem 12: (Basis of Eigenvectors)

If an $n \times n$ matrix A possesses n unique eigenvalues, it implies that A has a set of eigenvectors, $X_1, X_2, ..., X_n$ that forms a basis for the vector space $\mathbb{R}^n$.

Theorem 13: (Symmetric matrices)

An $n \times n$ symmetric matrix A possesses an orthonormal set of eigenvectors, $X_1, X_2, ...,$ X_n that forms a basis for the vector space $\mathbb{R}^n$.

Definition: (Similar Matrices, Similarity Transformation)

Two square matrices A and B, are said to be similar if there exists an invertible matrix P such that: $B = P^{-1}AP$. Here, P is a non-singular (invertible) matrix that facilitates the transformation. This transformation is called similarity transformation.

Theorem 14: (Eigenvalues and Eigenvectors of Similar Matrices)

If a matrix B is similar to A, then B has the same eigenvalues as A.

Furthermore, if X is an eigenvector of A, then $Y = P^{-1}X$ is an eigenvector of B corresponding to the same eigenvalue.

14.4 SOME OTHER PROPERTIES OF EIGEN VALUES

1. Sum of Eigenvalues (Trace);

The trace of a matrix A is equal to the sum of its eigenvalues.

$$trace(A) = \lambda_1 + \lambda_2 + ... \lambda_n$$

2. Product of Eigenvalues (Determinant):

The determinant of a matrix A is a equal to the product of its eigenvalues.

$$\det(A) = \lambda_1 \times \lambda_2 \times ... \lambda_n$$

3. Eigenvalues of Inverse:

If the eigenvalues of an invertible matrix A are $\lambda_1, \lambda_2, ..., \lambda_n$. Then the eigenvalues of matrix A^{-1} are $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, ..., \frac{1}{\lambda_n}$

For Example: The eigenvalues of $A = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$ are $\lambda_1 = 2, \lambda_2 = 3$.

However, the eigenvalues of $A^{-1} = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$ are $\frac{1}{2}$ and $\frac{1}{3}$.

4. Eigenvalues of Product:

The eigenvalues of the product AB are the same as the eigenvalues of BA, where A and B are square matrices.

5. Eigenvalues of Power:

If the eigenvalues of A are $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then the eigenvalues of matrix A^k (where k is a positive integer) are $\lambda_1{}^k, \lambda_2{}^k, \ldots, \lambda_n{}^k$.

14.5 DIAGONALIZATION OF A MATRIX:

Theorem 15: (Diagonalization of a Matrix)

If an $n \times n$ matrix A possesses a set of eigenvectors forming a basis of $\mathbb{R}^n$, then the matrix X with these eigenvectors as its columns diagonalizes A.

$$D = X^{-1} AX$$

Here, the diagonal matrix D has the eigenvectors of A along its main diagonal. And X is the matrix with eigenvectors of A as column vectors. Also,

$$D^m = X^{-1} A^m X \quad \text{for } m = 2, 3, \ldots.$$

Example 14.7: (Diagonalization of Matrix)

Consider a matrix $A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$ for which we will find the eigenvalues, eigenvectors, and then demonstrate its diagonalization.

1. Eigenvalues:

The eigenvalues are obtained by solving the characteristic equation det $(A - \lambda I) - 0$

$$\det\left(\begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix}\right) = (4 - \lambda)\,(3 - \lambda) - (2 \cdot 1) = \lambda^2 - 7\lambda + 10 - 0$$

This gives us two eigenvalues: $\lambda_1 = 2$ and $\lambda_2 = 5$.

2. Eigenvectors:

Now, for each eigenvalue, we find the corresponding eigenvector V by solving $(A - \lambda I)$ $V = 0$.

For $\lambda = 2$:

$$(A - 2I)V = \begin{bmatrix} 4-2 & 1 \\ 2 & 3-2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Solving this system of equations, we find an eigenvector $V_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$

For $\lambda = 5$:

$$(A - 2I)V = \begin{bmatrix} 4-5 & 1 \\ 2 & 3-5 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Solving this system of equations, we find an eigenvector $V_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

3. Diagonalization:

To diagonalize A we form the matrix P with the eigenvectors as columns and D as the diagonal matrix of eigenvalues. Therefore,

$$P = \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix} \text{and } D = \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}$$

Then, the diagonalization is given by $A = PDP^{-1}$

$$PDP^{-1} = \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} = A$$

Moreover, $$PD^2P^{-1} = \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}^2\begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 18 & 7 \\ 14 & 11 \end{bmatrix} = A^2$$

This demonstrates the diagonalization of the matrix A.

Example 14.8 (Inverse of a matrix using Kayley-Hamilton theorem)

Consider a matrix $A = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{bmatrix}$.

The characteristic equation is given by $\det(A - \lambda I) = 0$

$$\begin{bmatrix} 2-\lambda & -1 & 1 \\ -1 & 2-\lambda & -1 \\ 1 & -1 & 2-\lambda \end{bmatrix} = 0$$

$$(2-\lambda)\begin{vmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{vmatrix} + \begin{vmatrix} -1 & -1 \\ 1 & 2-\lambda \end{vmatrix} + \begin{vmatrix} -1 & 2-\lambda \\ 1 & -1 \end{vmatrix} = 0$$

$$(2 - \lambda)(\lambda^2 - 4\lambda + 3) + (\lambda - 1) + (\lambda - 1)$$

$$\lambda^2 - 6\lambda^2 + 9\lambda - 4 = 0$$

Corresponding, using Kayley-Hamilton theorem,

$$A^3 - 6A^2 + 9A - 4I = 0$$

Multiplying by A^{-1} in above equation we get,

$$A^2 - 6A + 9I - 4A^{-1} = 0$$

$$A^{-1} = \frac{1}{4}(A^2 - 6A + 9I)$$

$$A^{-1} = \frac{1}{4}\left(\begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{bmatrix}^2 - 6\begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{bmatrix} + 9\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right)$$

$$A^{-1} = \frac{1}{4}\left(\begin{bmatrix} 6 & -5 & -5 \\ -5 & -6 & -5 \\ 5 & -5 & 6 \end{bmatrix} + \begin{bmatrix} -12 & 6 & -6 \\ 6 & -12 & 6 \\ -6 & 6 & -12 \end{bmatrix} + \begin{bmatrix} 9 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 9 \end{bmatrix}\right)$$

$$A^{-1} = \frac{1}{4}\begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & 1 \\ -1 & 1 & 3 \end{bmatrix}$$

PROBLEMS

1. Determine A^2, A^3, and A^{-1}, where $A = \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$

2. Determine a real matrix $A = B^3$, where $B = \begin{bmatrix} 4 & 6 & 0 \\ 0 & 4 & 6 \\ 0 & 0 & 4 \end{bmatrix}$

3. Find a polynomial having the following matrix as a root:

(a) $A = \begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$,

(b) $A = \begin{bmatrix} 1 & 3 & 4 \\ 1 & 4 & 6 \\ 0 & 1 & 2 \end{bmatrix}$

4. For each of the following matrices, find the eigenvalues and corresponding eigenvectors:

(a) $A = \begin{bmatrix} 6 & -1 \\ 2 & 3 \end{bmatrix}$ (b)$A = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 2 \end{bmatrix}$

5. Consider a matrix $A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$, then

(a) Find a non-singular matrix P such that $D = P^{-1} AP$ is diagonal.

(b) Find a matrix B such that $B^2 = A$.

(c) Find A^4 and $f(A)$, where $f(u) = 2u^2 - 3u + 5$.

6. Find a real 2×2 symmetric matrix A with eigenvalues:

(a) $\lambda_1 = 1, \lambda_2 = 4$ and eigenvector corresponding to λ_1 is given as $X = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

(b) $\lambda_1 = 3, \lambda_2 = 2$ and eigenvector corresponding to λ_2 is given as $X = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

7. Let A be a square matrix with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$. Prove that the sum of the eigenvalues (trace) is equal to the sum of the elements on the main diagonal, and the product of the eigenvalues (determinant) is equal to the determinant of matrix A.

8. If A is a square matrix with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$ show that the transpose of A has the same eigenvalues.

9. If A is invertible with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$. Then find the eigenvalues of the inverse of A.

10. If A is a square matrix with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$. Then find the eigenvalues of matrix cA, where c is a scalar.

11. Let A and B be matrices with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$ and $\mu_1, \mu_2, ..., \mu_n$, respectively. Then find the eigenvalues of $A + B$.

12. A and B be matrices with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$ and $\mu_1, \mu_2, ..., \mu_n$, respectively. Then find the eigenvalues of AB.

13. Determine the Inverse of a matrix $A = \begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ -2 & 2 & 1 \end{bmatrix}$ using Kayley-Hamilton theorem.

◆

APPENDIX-A

Table - **I**: Cululative Binomial Probabilites $P(X \leq x)$

		P										
n	x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
1	0	0.9000	0.8000	0.7000	0.6000	0.5000	0.4000	0.3000	0.2000	0.1000	0.0500	0.0100
2	0	0.8100	0.6400	0.4900	0.3600	0.2500	0.1600	0.0900	0.0400	0.0100	0.0025	0.0001
	1	0.9900	0.9600	0.9100	0.8400	0.7500	0.6400	0.5100	0.3600	0.1900	0.0975	0.0199
3	0	0.7290	0.5120	0.3430	0.2160	0.1250	0.0640	0.0270	0.0080	0.0010	0.0001	0.0000
	1	0.9720	0.8960	0.7840	0.6480	0.5000	0.3520	0.2160	0.1040	0.0280	0.0073	0.0003
	2	0.9990	0.9920	0.9730	0.9360	0.8750	0.7840	0.6570	0.4880	0.2710	0.1426	0.0297
4	0	0.6561	0.4096	0.2401	0.1296	0.0625	0.0256	0.0081	0.0016	0.0001	0.0000	0.0000
	1	0.9477	0.8192	0.6517	0.4752	0.3125	0.1792	0.0837	0.0272	0.0037	0.0005	0.0000
	2	0.9963	0.9728	0.9163	0.8208	0.6875	0.5248	0.3483	0.1808	0.0523	0.0140	0.0006
	3	0.9999	0.9984	0.9919	0.9744	0.9375	0.8704	0.7599	0.5904	0.3439	0.1855	0.0394
5	0	0.5905	0.3277	0.1681	0.0778	0.0313	0.0120	0.0024	0.0003	0.0000	0.0000	0.0000
	1	0.9185	0.7373	0.5282	0.3370	0.1875	0.0870	0.0308	0.0067	0.0005	0.0000	0.0000
	2	0.9914	0.9421	0.8369	0.6826	0.5000	0.3174	0.1631	0.0579	0.0086	0.0012	0.0000
	3	0.9995	0.9933	0.9692	0.9130	0.8125	0.6630	0.4718	0.2627	0.0815	0.0226	0.0010
	4	1.0000	0.9997	0.9976	0.9898	0.6988	0.9222	0.8319	0.6723	0.4095	0.2262	0.0490
6	0	0.5314	0.2621	0.1176	0.0467	0.0156	0.0041	0.0007	0.0001	0.0000	0.0000	0.0000
	1	0.8857	0.6554	0.4202	0.2333	0.1094	0.0410	0.0109	0.0016	0.0001	0.0000	0.0000
	2	0.9842	0.9011	0.7443	0.5443	0.3438	0.1792	0.0705	0.0170	0.0013	0.0001	0.0000
	3	0.9987	0.9830	0.9295	0.8208	0.6563	0.4557	0.2557	0.0989	0.0159	0.0022	0.0000
	4	0.9999	0.9984	0.9891	0.9590	0.9806	0.7667	0.5798	0.3446	0.1143	0.0328	0.0015
	5	1.0000	0.9999	0.9993	0.9959	0.9844	0.9533	0.8824	0.7379	0.4686	0.2649	0.0585
7	0	0.4783	0.2097	0.0824	0.0280	0.0078	0.0016	0.0002	0.0000	0.0000	0.0000	0.0000
	1	0.8503	0.5767	0.3294	0.1586	0.0625	0.0188	0.0038	0.0004	0.0000	0.0000	0.0000
	2	0.9743	0.8520	0.6471	0.4199	0.2266	0.096	0.0288	0.0047	0.0002	0.0000	0.0000
	3	0.9973	0.9667	0.8740	0.7102	0.5000	0.2898	0.1260	0.0333	0.0027	0.0002	0.0000
	4	0.9998	0.9953	9.9712	0.9037	0.7734	0.5801	0.3529	0.1480	0.0257	0.0038	0.0000
	5	1.000	0.9996	0.9962	0.9812	0.9375	0.8414	0.6706	0.4233	0.1497	0.0444	0.0020
	6	1.000	1.0000	0.9998	0.9984	0.9922	0.9720	0.9176	0.7903	0.5217	0.301	0.0679
8	0	0.4305	0.1678	0.0576	0.0168	0.0039	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
	1	0.81311	0.5033	0.2553	0.1064	0.0352	0.0085	0.0013	0.0001	0.0000	0.0000	0.0000
	2	0.9619	0.7969	0.5518	0.3154	0.1445	0.0498	0.0113	0.0012	0.0000	0.0000	0.0000
	3	0.9950	0.9437	0.8059	0.5941	0.3633	0.1737	0.0580	0.0104	0.0004	0.0000	0.0000
	4	0.9996	0.9896	0.9420	0.8263	0.6367	0.4059	0.1941	0.0563	0.0050	0.0004	0.0000
	5	1.0000	0.9988	0.9887	0.9502	0.8555	0.6846	0.4482	0.2031	0.0381	0.0058	0.0001
	6	1.0000	0.9999	0.9887	0.9915	0.9648	0.8936	0.7447	0.4967	0.1869	0.0572	0.0027
	7	1.0000	1.0000	0.9999	0.9993	0.9961	0.9832	0.9424	0.8322	0.5695	0.3366	0.0773
9	0	0.3874	0.1342	0.0404	0.0101	0.0020	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.7748	0.4362	0.1960	0.0705	0.0195	0.0038	0.0004	0.0000	0.0000	0.0000	0.0000
	2	0.9470	0.7382	0.4628	0.2318	0.0889	0.0250	0.0043	0.0003	0.0000	0.0000	0.0000
	3	0.9917	0.9144	0.7297	0.4826	0.2539	0.0994	0.0253	0.003	0.0001	0.0000	0.0000
	4	0.9991	0.9804	0.9012	0.7334	0.5000	0.2666	0.0988	0.0196	0.0009	0.0000	0.0000
	5	0.9999	0.9969	0.9747	0.9006	0.7461	0.5174	0.2703	0.0856	0.0083	0.0006	0.0000
	6	1.0000	0.9997	0.9957	0.9750	0.9102	0.7682	0.5372	0.2618	0.0530	0.0084	0.0001
	7	1.0000	1.0000	0.9996	0.9962	0.9805	0.9295	0.8040	0.5638	0.2252	0.0712	0.0034
	8	1.0000	1.0000	1.0000	0.9997	0.9980	0.9899	0.9596	0.8658	0.6126	0.3698	0.0865

Table - I: Culalative Binomial Probabilites $P(X \leq x)$ *(continued)*

		P										
n	*x*	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
10	0	0.3487	0.1074	0.0282	0.0060	0.0010	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.7361	0.3758	0.1493	0.0464	0.0107	0.0017	0.0001	0.0000	0.0000	0.0000	0.0000
	2	0.9298	0.6778	0.3828	0.1673	0.0547	0.0123	0.0016	0.0001	0.0000	0.0000	0.0000
	3	0.9872	0.8791	0.6496	0.3823	0.1719	0.0548	0.0106	0.0009	0.0000	0.0000	0.0000
	4	0.9984	0.9672	0.8497	0.6331	0.3770	0.1662	0.0473	0.0064	0.0001	0.0000	0.0000
	5	0.9999	0.9936	0.9527	0.8338	0.6230	0.3669	0.1503	0.0328	0.0016	0.0001	0.0000
	6	1.0000	0.9991	0.9894	0.9452	0.8281	0.6177	0.3504	0.1209	0.0128	0.0010	0.0000
	7	1.0000	0.9999	0.9984	0.9877	0.9453	0.8327	0.6172	0.3222	0.0702	0.0115	0.0001
	8	1.0000	1.0000	0.9999	0.9983	0.989	0.9536	0.8507	0.6242	0.2639	0.0861	0.0043
	9	1.0000	1.0000	1.0000	0.9999	0.9990	0.9940	0.9718	0.8926	0.6513	0.4013	0.0956
11	0	0.3138	0.0859	0.0198	0.0036	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.6974	0.3221	0.1130	0.0302	0.0059	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9104	0.6174	0.3127	0.1189	0.0327	0.0059	0.0006	0.0000	0.0000	0.0000	0.0000
	3	0.9815	0.8389	0.5696	0.2963	0.1133	0.0293	0.0043	0.0002	0.0000	0.0000	0.0000
	4	0.9972	0.9496	0.7897	0.5328	0.2744	0.0994	0.0216	0.0020	0.0000	0.0000	0.0000
	5	0.9997	0.9883	0.9218	0.7535	0.5000	0.2465	0.0782	0.0117	0.0003	0.0000	0.0000
	6	1.0000	0.9980	0.9784	0.9006	0.7256	0.4672	0.2103	0.0504	0.0028	0.0001	0.0000
	7	1.0000	0.9998	0.9957	0.9707	0.8867	0.7037	0.4304	0.1611	0.0185	0.0016	0.0000
	8	1.0000	1.0000	0.9994	0.994	0.9674	0.8811	0.6873	0.3826	0.0896	0.0152	0.0002
	9	1.0000	1.0000	1.0000	0.9993	0.9941	0.9698	0.8870	0.6779	0.3026	0.1019	0.0052
	10	1.0000	1.0000	1.0000	1.0000	0.9995	0.9964	0.9802	0.9141	0.6862	0.4312	0.1047
12	0	0.2824	0.0687	0.0138	0.0022	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.6590	0.2749	0.0850	0.0196	0.0032	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.8891	0.5583	0.2528	0.0834	0.0193	0.0028	0.0002	0.0000	0.0000	0.0000	0.0000
	3	0.9744	0.7946	0.4925	0.2253	0.0730	0.0153	0.0017	0.0001	0.0000	0.0000	0.0000
	4	0.9957	0.9274	0.7237	0.4382	0.1938	0.0573	0.0095	0.0006	0.0000	0.0000	0.0000
	5	0.9995	0.9806	0.8822	0.6652	0.3872	0.1582	0.0386	0.0039	0.0001	0.0000	0.0000
	6	0.9999	0.9961	0.9614	0.8418	0.6128	0.3348	0.1178	0.0194	0.0005	0.0000	0.0000
	7	1.0000	0.9994	0.9905	0.9427	0.8062	0.5618	0.2763	0.0726	0.0043	0.0002	0.0000
	8	1.0000	0.9999	0.9983	0.9847	0.9270	0.7747	0.5075	0.2054	0.0256	0.0022	0.0000
	9	1.0000	1.0000	0.9998	0.9972	0.9807	0.9166	0.7472	0.4417	0.1109	0.0196	0.0002
	10	1.0000	1.0000	1.0000	0.9997	0.9968	0.9804	0.9150	0.7251	0.3410	0.1184	0.0062
	11	1.0000	1.0000	1.0000	1.0000	0.9998	0.9978	0.9862	0.9313	0.7176	0.4596	0.1136
13	0	0.2542	0.00550	0.0097	0.0013	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.6213	0.2336	0.0637	0.0126	0.0017	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.8661	0.5017	0.2025	0.0579	0.0112	0.0013	0.0001	0.0000	0.0000	0.0000	0.0000
	3	0.9658	0.7473	0.4206	0.1686	0.046	0.0078	0.0007	0.0000	0.0000	0.0000	0.0000
	4	0.9935	0.9700	0.346	0.5744	0.2905	0.0977	0.0182	0.0012	0.0000	0.0000	0.0000
	5	0.9991	0.9700	0.8346	0.5744	0.2905	0.0977	0.0182	0.0012	0.0000	0.0000	0.0000
	6	0.9999	0.9930	0.9376	0.7712	0.5000	0.2288	0.0624	0.0070	0.0001	0.0000	0.0000
	7	1.0000	0.9958	0.9818	0.9023	0.7095	0.4256	0.1654	0.0300	0.0009	0.0000	0.0000
	8	1.0000	0.9988	0.9960	0.9679	0.8666	0.6470	0.3457	0.0991	0.0065	0.0003	0.0000
	9	1.0000	1.0000	0.9993	0.9922	0.9539	0.314	0.5794	0.2527	0.0342	0.0031	0.0000
	10	1.0000	1.0000	0.9999	0.9987	0.9888	0.9421	0.7975	0.4983	0.1339	0.0245	0.0003
	11	1.0000	1.0000	1.0000	0.9999	0.9983	0.9874	0.9363	0.7664	0.3787	0.1354	0.0072
	12	1.0000	1.0000	1.0000	1.0000	0.9999	0.9987	0.9903	0.9450	0.7458	0.4867	0.1225

Table - I: Cululative Binomial Probabilites $P(X \leq x)$ (*continued*)

		P										
n	*x*	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
14	0	0.2288	0.0440	0.0068	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.5846	0.1979	0.0475	0.0081	0.0009	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.8416	0.4481	0.1608	0.0398	0.0398	0.0065	0.0006	0.0000	0.0000	0.0000	0.0000
	3	0.9559	0.6982	0.3552	0.1243	0.0287	0.0039	0.0002	0.0000	0.0000	0.0000	0.0000
	4	0.9908	0.8702	0.5842	0.2793	0.0898	0.0175	0.0017	0.0000	0.0000	0.0000	0.0000
	5	0.9985	0.9561	0.7805	0.4859	0.2120	0.0583	0.0083	0.0004	0.0000	0.0000	0.0000
	6	0.9998	0.9884	0.9067	0.6925	0.3953	0.1501	0.0315	0.0024	0.0000	0.0000	0.0000
	7	1.0000	0.9976	0.9685	0.8499	0.6047	0.3075	0.0933	0.0116	0.0002	0.0000	0.0000
	8	1.0000	0.9996	0.9917	0.9417	0.7880	0.5141	0.2195	0.0439	0.0015	0.0000	0.0000
	9	1.0000	1.0000	0.9983	0.9825	0.9102	0.7207	0.4158	0.1298	0.0092	0.0004	0.0000
	10	1.0000	1.0000	0.9998	0.9961	0.9713	0.8757	0.6448	0.3018	0.0441	0.0042	0.0000
	11	1.0000	1.0000	1.0000	0.9994	0.9935	0.9602	0.8392	0.5519	0.1584	0.0301	0.0003
	12	1.0000	1.0000	1.0000	0.9999	0.9991	0.9919	0.9525	0.8021	0.4154	0.1530	0.0084
	13	1.0000	1.0000	1.0000	1.0000	0.9999	0.9992	0.9932	0.950	0.7712	0.5123	0.1313
15	0	0.2059	0.0352	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.5490	0.1671	0.0353	0.0052	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.8159	0.3980	0.1268	0.0271	0.0037	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9444	0.6482	0.2969	0.0905	0.0176	0.0019	0.0001	0.0000	0.0000	0.0000	0.0000
	4	0.9873	0.8358	0.5155	0.2173	0.0592	0.0093	0.0007	0.0000	0.0000	0.0000	0.0000
	5	0.9978	0.9389	0.7216	0.4032	0.1509	0.0338	0.0037	0.0001	0.0000	0.0000	0.0000
	6	0.9997	0.9819	0.8689	0.6098	0.3036	0.0950	0.0152	0.0008	0.0000	0.0000	0.0000
	7	1.0000	0.9958	0.9500	0.7869	0.5000	0.2131	0.0500	0.0042	0.0000	0.0000	0.0000
	8	1.0000	0.9992	0.9848	0.9050	0.6964	0.3902	0.1311	0.0181	0.0003	0.0000	0.0000
	9	1.0000	0.9999	0.9963	0.9662	0.8492	0.5968	0.2784	0.0611	0.0022	0.0001	0.0000
	10	1.0000	1.0000	0.9993	0.9907	0.9408	0.7827	0.4845	0.1642	0.0127	0.0006	0.0000
	11	1.0000	1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.3518	0.0556	0.0055	0.0000
	12	1.0000	1.0000	1.0000	0.9997	0.9963	0.9729	0.8732	0.6020	0.8141	0.0362	0.0004
	13	1.0000	1.0000	1.0000	1.0000	0.9995	0.9948	0.9647	0.8329	0.4510	0.1710	0.0096
	14	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9953	0.9648	0.7941	0.5367	0.1399
20	0	0.1216	0.0115	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.3917	0.0692	0.0076	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.6769	0.2061	0.0355	0.0036	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.8670	0.4114	0.1071	0.0160	0.0013	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.9568	0.6296	0.2375	0.0510	0.0059	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.9887	0.8042	0.4164	0.1256	0.0207	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.9976	0.9133	0.6080	0.2500	0.0577	0.0065	0.0003	0.0000	0.0000	0.0000	0.0000
	7	0.9996	0.9679	0.7723	0.4159	0.1316	0.0210	0.0013	0.0000	0.0000	0.0000	0.0000
	8	0.9999	0.9900	0.8867	0.5956	0.2517	0.0565	0.0051	0.0001	0.0000	0.0000	0.0000
	9	1.0000	0.9974	0.9520	0.7553	0.4119	0.1275	0.0171	0.0006	0.0000	0.0000	0.0000
	10	1.0000	0.9994	0.9829	0.8725	0.5881	0.2447	0.0480	0.0026	0.0000	0.0000	0.0000
	11	1.0000	0.9999	0.9949	0.9435	0.7483	0.4044	0.1133	0.0100	0.0001	0.0000	0.0000
	12	1.0000	1.0000	0.9987	0.9790	0.8684	0.5841	0.2277	0.0321	0.0004	0.0000	0.0000
	13	1.0000	1.0000	0.9997	0.9935	0.9423	0.7500	0.3920	0.0867	0.0024	0.0000	0.0000
	14	1.0000	1.0000	1.0000	0.9984	0.9793	0.8744	0.5836	0.1958	0.0113	0.0003	0.0000
	15	1.0000	1.0000	1.0000	0.9997	0.9941	0.9490	0.725	0.3704	0.0432	0.0026	0.0000
	16	1.0000	1.0000	1.0000	1.0000	0.9987	0.9840	0.8929	0.5886	0.1330	0.0159	0.0000
	17	1.0000	1.0000	1.0000	1.0000	0.9998	0.9964	0.9645	0.7939	0.3231	0.0755	0.0010
	18	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9924	0.9308	0.6083	0.2642	0.0169
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992	0.9885	0.8784	0.6415	0.1821

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

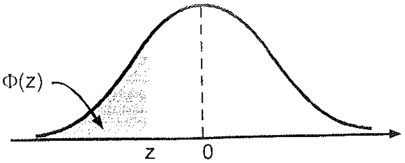


Table II: Cumulative Standard Normal Distribution

z	–0.09	–0.08	–0.07	–0.06	–0.05	–0.04	–0.03	–0.02	–0.01	–0.00
–3.9	0.000033	0.000034	0.000036	0.000037	0.000039	0.000041	0.000042	0.000044	0.000046	0.000048
–3.8	0.000050	0.000052	0.000054	0.000057	0.000059	0.000062	0.000064	0.000067	0.000069	0.000072
–3.7	0.000075	0.000078	0.000082	0.000085	0.000088	0.000092	0.000096	0.000100	0.000104	0.000108
–3.6	0.000112	0.000117	0.000121	0.000126	0.000131	0.000136	0.000142	0.000147	0.000153	0.000159
–3.5	0.000165	0.000172	0.000179	0.000185	0.000193	0.000200	0.000208	0.000216	0.000224	0.000233
–3.4	0.000242	0.000251	0.000260	0.000270	0.000280	0.000291	0.000302	0.000313	0.000325	0.000337
–3.3	0.000350	0.000362	0.000376	0.000390	0.000404	0.000419	0.000434	0.000450	0.000467	0.000483
–3.2	0.000501	0.000519	0.000538	0.000557	0.000577	0.000598	0.000619	0.000641	0.000664	0.000687
–3.1	0.000711	0.000736	0.000762	0.000789	0.000816	0.000845	0.000874	0.000904	0.000935	0.000968
–3.0	0.001001	0.001035	0.001070	0.001107	0.001144	0.001183	0.001223	0.001264	0.001306	0.001350
–2.9	0.001395	0.001441	0.001489	0.001538	0.001589	0.001641	0.001695	0.001750	0.001807	0.001866
–2.8	0.001926	0.001988	0.002052	0.002118	0.002186	0.002256	0.002327	0.002401	0.002477	0.002555
–2.7	0.002635	0.002718	0.002803	0.002890	0.002980	0.003072	0.003167	0.003264	0.003364	0.003467
–2.6	0.003573	0.003681	0.003793	0.003907	0.004025	0.004145	0.004269	0.004396	0.004527	0.004661
–2.5	0.004799	0.004940	0.005085	0.005234	0.005386	0.005543	0.005703	0.005868	0.006037	0.006210
–2.4	0.006387	0.005669	0.006756	0.006947	0.007143	0.007344	0.007549	0.007760	0.007976	0.008198]
–2.3	0.008424	0.008656	0.008894	0.009137	0.009387	0.009642	0.009903	0.010170	0.010444	0.010724
–2.2	0.011011	0.011304	0.011604	0.011911	0.012224	0.012545	0.012874	0.013209	0.013553	0.013903
–2.1	0.014262	0.014629	0.015003	0.015386	0.015778	0.016177	0.016586	0.017003	0.017429	0.017864
–2.0	0.018309	0.018763	0.019226	0.019699	0.020182	0.020675	0.021178	0.021692	0.022216	0.022750
–1.9	0.023295	0.023852	0.024419	0.024998	0.025588	0.026190	0.026803	0.027429	0.028067	0.028717
–1.8	0.029379	0.030054	0.030742	0.031443	0.032157	0.032884	00.33625	0.034379	0.035148	0.035930
–1.7	0.036727	0.037538	0.038364	0.039204	0.040059	0.040929	0.041815	0.042716	0.043633	0.044656
–1.6	0.045514	0.046479	0.047460	0.048457	0.049471	0.050503	0.051551	0.052616	0.053699	0.054799
–1.5	0.055917	0.057053	0.058208	0.059380	0.060571	0.061780	0.0631008	0.064256	0.065522	0.066807
–1.4	0.068112	0.069437	0.070781	0.072145	0.073529	0.074934	0.076359	0.077804	0.079270	0.080757
–1.3	0.082264	0.083793	0.085343	0.086915	0.088508	0.090123	0.091759	0.093418	0.095098	0.096801
–1.2	0.098525	0.100273	0.102042	0.103835	0.105650	0107488	0.109349	0.111233	0.113140	0.115070
–1.1	0.117023	0.119000	0.121001	0.123024	0.125072	0.127143	0.129238	0.111233	0.113140	0.115070
–1.0	0.137857	0.140071	0.142310	0.144572	0.146859	0.149170	0.151505	0.153864	0.156248	0.158655
–0.9	0.161087	0.163543	0.166023	0.168528	0.171056	0.173609	0.176185	0.178786	0.181411	0.184060
–0.8	0.186733	0.189430	0.192150	0.194894	0.197662	0.200454	0.203269	0.206108	0.208970	0.211855
–0.7	0.214764	0.217695	0.220650	0.223627	0.226627	0.229650	0.232695	0.0235762	0.238852	0.241964
–0.6	0.245097	0.248252	0.251429	0.245627	0.257846	0.261086	0.264347	0.267629	0.270931	0.274253
–0.5	0.277595	0.280957	0.284339	0.287740	0.291160	0.294599	0.298056	0.301532	0.305026	0.308538
–0.4	0.312067	0.315614	0.319178	0.322758	0.326355	0.329969	0.333598	0.337243	0.340903	0.344578
–0.3	0.348268	0.351973	0.355691	0.359424	0.363169	0.366928	0.370700	0.374484	0.378281	0.382089
–0.2	0.385908	0.389739	0.393580	0.397432	0.401294	0.405165	0.409046	0.412936	0.416834	0.420740
–0.1	0.424655	0.428576	0.432505	0.436441	0.440382	0.444330	0.448283	0.452242	0.456205	0.460172
–0.0	0.464144	0.468119	0.472097	0.476078	0.480061	0.484047	0.488033	0.492022	0.496011	0.500000

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\,du$$

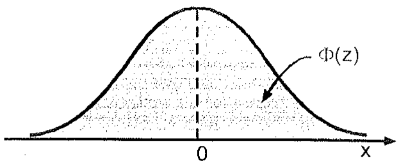


Table II: Cumulative Standard Normal Distribution (*continued*)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.751748	0.754903
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.80106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.9369992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0..992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.99457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	9.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	9.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

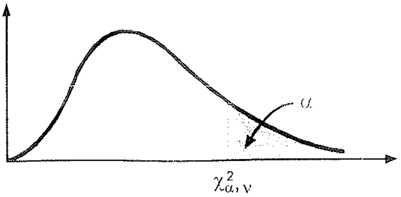


Table III Percentage Points $\chi^2_{\alpha,\nu}$ of the Chi-Squared Distribution

ν \ α	.995	.990	.975	.950	.900	.500	.100	.050	.025	.010	.005
1	.00+	.00+	.00+	.00+	.02	.45	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	1.39	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	2.37	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.,34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	3889	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	53.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

ν = degree of freedom.

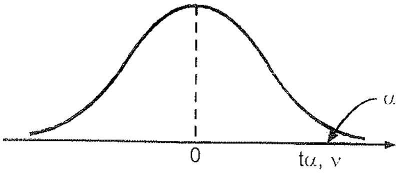


Table IV: Percentage Points $t_{\alpha,\nu}$ of the t Distribution

ν \ α	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.474	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	.1345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.098	2.539	2.861	3.147	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.08	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

ν = degrees of freedom.

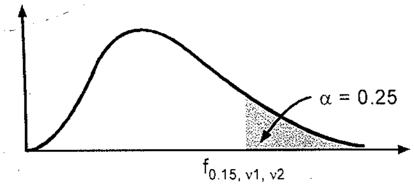


Table V: Percentage Points f_{α,v_1,v_2}, of the F Distribution

Degree of freedom for the denominator (v^2)

v_2 \ v_1	Degrees of freedom for the numerator (v_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	5.83	7.50	8.20	8.58	8.882	8.98	9.10	9.19	9.26	9.32	9.41	9.49	9.58	9.63	9.67	9.71	9.76	9.80	9.85
2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.48
3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47
4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.88	1.87	1.87	1.87	1.87
6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74
7	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.70	1.69	1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65
8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.62	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.54	1.53	1.53
10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.48
11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.45
12	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.43	1.42
13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40
14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.39	1.38
15	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.37	1.36
16	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
17	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33
18	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.32
19	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.30
20	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.29
21	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28
22	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.37	1.36	1.34	1.33	1.32	1.31	1.30	1.29	1.28
23	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28	1.27
24	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28	1.26
25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.27	1.25
26	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.35	1.34	1.32	1.31	1.3o	1.29	1.28	1.26	1.25
27	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.24
28	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24
29	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.23
30	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23
40	1,35	1,44	1,42	1,49	1,38	1,36	1,35	1,35	1,34	1,33	1,31	1,39	1.28	1.26	1.25	1.24	1.22	1.21	1.19
60	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.27	1.25	1.24	1.22	1.21	1.19	1.17	1.15
120	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.10
∞	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.00

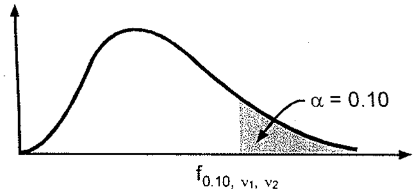


Table V: Percentage Points $f_{\alpha,\nu_1,\nu_2,}$ of the F Distribution (*continued*)

Degree of freedom for the denominator (v^2)

ν_2 / ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.3	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.3o	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3. 23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	.252	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.29	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.8	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2,83	2,54	2,33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.6	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.56	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.03	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	1.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

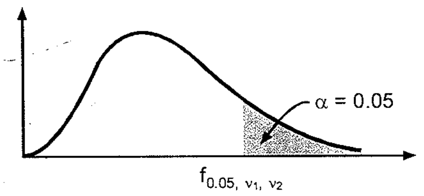


Table V: Percentage Points f_{α, v_1, v_2}, of the F Distribution (*continued*)

Degree of freedom for the denominator (v^2)

v_2 / v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.45	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.33	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.60	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	.207
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.53	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	.207	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.1	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.58	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.56	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.55	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.2	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

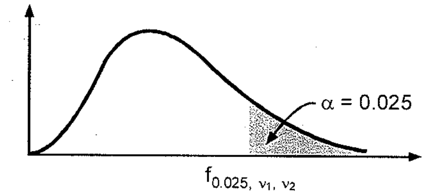


Table V: Percentage Points $f_{\alpha,v_1,v_2,}$ of the F Distribution (*Continued*)

Degree of freedom for the denominator (v^2)

v_2 \ v_1	Degrees of freedom for the numerator (v_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	0	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	2.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	482	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.8	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.35
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.921	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	233	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.00	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.08	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.3	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.80	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

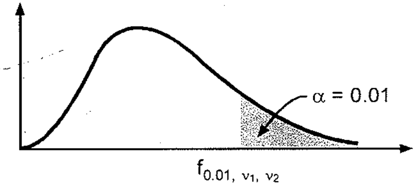


Table VI: Percentage Points $f_{\alpha,v_1,v_2,}$ of the F Distribution (*Continued*)

Degree of freedom for the denominator (v^2)

v_2 \ v_1	Degrees of freedom for the numerator (v_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	0	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	58.50	99.0	9.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.7	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.46
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.242	4.89	4.36	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.7	2.58	2.59
20	8.18	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

BIBLIOGRAPHY

F. A. Graybill (1961), An introduction to linear statistical models-Vol-1 McGraw-Hill, New York.

NPTEL Lecture notes on Probability and Statistical Inference by Prof. Shalabh, Department of Mathematics and Statistics, IIT Kanpur, India.

D. C. Montgomery and G. C. Runger (2018), Applied Statistics and Probability for Engineers 3rd ed., John Wiley & Sons, New York.

W. W. Hines and D. C. Montgomery (1990), Probability and Statistics in Engineering and Management Science, 3rd ed., John Wiley & Sons, New York.

S. Ross, (1987), Introduction to Probability and Statistics for Engineers and Scientists, John Wiley & Sons, New York.

C. Daniel and F. Wood (1980), Fitting Equations to Data, 2nd ed., John Wiley & Sons, New York

N. Draper and H. Smith (1998), Applied Regression Analysis, 3rd ed., John Wiley & Sons, New York.

R. H. Myers (1990), Classical and Modern Regression with Applications, 2nd ed., PWS-Kent, Boston.

D. C. Montgomery, E. A. Peck, and G. G. Vining (2001), Introduction to Linear Regression Analysis, 3rd ed., John Wiley & Sons, New York.

S. Lipschutz and M. L. Lipson (2009), Linear Algebra. 4th ed., Schaum's Outline Series, Mc Graw Hill, New Delhi.

MODEL QUESTION PAPER

STATISTICS, STATISTICAL MODELLING AND DATA ANALYTICS

End Term Examination, Sixth Semester [B. Tech.]

Time: 3 Hrs. *Max. Marks: 75*

Note: Attempt five questions in all including Q. No. 1 which is compulsory. Select one question from each unit.

Q.1. (*a*) What do you mean by statistical inference?

(*b*) Give the assumptions of simple linear regression model.

(*c*) State and prove the Triangle inequality.

(*d*) Write any three properties of Eigen values.

(*e*) Discuss the role of Cauchy sequence concept in calculus and analysis

(3×5=15)

UNIT-I

Q.2. (*a*) If two independent random sample of size $n_1 = 7$ and $n_2 = 13$ are taken from a normal population what is the probability that the variance of the first sample will be at least three times as large that of the second sample? **(7)**

(*b*) A botanist has produced a new variety of hybrid soy plant that is better able to withstand drought than other varieties. The botanist knows the seed germination for the parent plants is 75%, but does not know the seed germination for the new hybrid. He tests the claim that it is different from the parent plants. To test this claim, 450 seeds from the hybrid plant are tested and 321 have germinated. Use a 5% level of significance to test this claim that the germination rate is different from 75%. **(8)**

Q.3. (*a*) Let X and Y be jointly continuous random variables with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} cx+1 & x,y \geq 0, x+y<1 \\ 0 & \text{Otherwise} \end{cases}$$

Show the range of (X,Y), R_{XY}, in the $x-y$ plane.

Find the constant c.

Find the marginal PDFs $f_X(x)$ and $f_Y(y)$

Find $P(y < 2x^2)$ **(8)**

(*b*) Consider the following system of linear equations:

$$2x + y - z = 5$$
$$4x - 2y + 3z = 3$$
$$-2x + 3y + 2z = 8$$

Write down the augmented matrix and perform Gaussian elimination to bring the system to upper triangular form. Then, use back substitution to find the values of x, y and z. **(7)**

UNIT-II

Q.4. (*a*) State and Prove Gauss-Markov theorem. **(7)**

(*b*) Consider three different teaching methods (A, B, and C) and their corresponding test scores for a sample of students:

Method A: [75, 82, 78, 88, 92]

Method B: [68, 75, 80, 85, 88]

Method C: [72, 78, 84, 90, 94]

Perform a one-way ANOVA to test whether there are any significant differences in the mean test scores among the three teaching methods. Provide the ANOVA table and state your conclusion. **(8)**

Q.5. (*a*) Consider the following data for a multiple linear regression model with two independent variables X_1 and X_2 and a dependent variable Y:

X_1	X_2	Y
2	4	20
3	5	25
4	6	30
5	7	35

(*i*) Fit a multiple linear regression model and write down the estimated regression equation.

(*ii*) Calculate the sum of squared residuals (SSE) for the model.

(*iii*) Calculate the coefficient of determination R^2 for the model. **(8)**

(*b*) Explain Residual analysis and the various methods for scaling residuals. **(7)**

UNIT-III

Q.6. (*a*) Let X be a metric space.

(*i*) Suppose that for some $\varepsilon > 0$ that every ε-neighborhood of every point has compact closure. Prove that X is complete. **(5)**

(*ii*) Suppose that for each $x \in X$ there is an $\varepsilon > 0$ that $N_\varepsilon(x)$ has compact closure. Show by example that X is not necessarily complete. **(5)**

(*b*) Check the convergence of the following sequences:

(*i*) $\lim_{n\to\infty} \frac{1}{n}$ (*ii*) $\lim_{n\to\infty} n\sin\left(\frac{1}{n}\right)$ (*iii*) $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$ **(5)**

Q.7. (*a*) Prove that:

(*i*) Set of $\mathbb{Q}$ rational numbers is countably infinite and set of $\mathbb{R}$ real numbers is uncountable.

(*ii*) Every compact subset of a metric space is closed and bounded. **(8)**

(*b*) If x, y and z are points on a metric space (X, d), show that

$d(x, y) \geq |d(x, z) - d(y, z)|$ **(7)**

UNIT-IV

Q.8. (*a*) Find Eigen values and Eigen vectors of the matrix $A = \begin{bmatrix} 8 & -6 & 2 \\ -6 & 7 & -4 \\ 2 & -4 & 3 \end{bmatrix}$. **(8)**

(b) Determine the Inverse of a matrix $A = \begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ -2 & 2 & 1 \end{bmatrix}$ using Kayley-Hamilton theorem. **(7)**

Q.9. (*a*) If A is invertible with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then find the eigenvalues of the inverse of A. **(8)**

(*b*) Check whether the vectors: $U_1 = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$ $U_2 = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix}$, and U_3 = are linearly independent. **(7)**

❑❑❑

INDEX

A Left-sided Test, 88
A Right-sided Test, 88
A Two-sided Test, 87
Akaike's Information Criterion (AIC), 248
Analysis of Covariance, 203
  - Single Factor ANOVA, 183
Analytical Methods for Selecting a Transformation on Study Variable (The Box-Cox Method), 230
ANOVA for a Two-way Factorial Design, 198
Assumptions Required for the Analysis of Single Factor ANOVA, 184

B...

Bar Charts and Histograms, 12
Basic Set Theory, 265
Basis and Dimensions, 291
Bayesian Information Criterion (BIC), 249
Binomial Distribution, 29
Boxplots, 13

C...

Cauchy Sequence, 272
Cayley-Hamilton theorem, Characteristic Polynomial, 295
Central Limit Theorem, 53
Checking of the Linear Relationship Between Study and Explanatory Variables, 211
Closed Set, 269
Coefficient of Determination, 160, 244, 245
Compact Sets, 272
Completeness in metric spaces, 274
Components of a Formal Hypothesis Test, 86
  - Metric Space, 268
Computational Techniques for Variable Selection, 251
Confidence Interval Interpretations, 124
Confidence Interval Summary, 82
Confidence Intervals about the Mean (μ) when the Population Standard Deviation (σ) is Known, 77
Confidence Intervals about the Mean (μ) when the Population Standard Deviation (σ) is Unknown, 79
Confidence Intervals about the Population Proportion (p), 81
Connected Metric Space, 276
Construct and Interpret a Confidence Interval About the Difference of Two Independent Means, 115
Construct and Interpret a Confidence Interval about the Differences of the Data for Matched Pairs, 124
Continuous Random Variables, 21
Convergence, 271
Covar, 171

D...

Descriptive Measures, 3
  - Statistics, 2
Diagnostic Plots, 217
Diagonalization of a Matrix, 302
Directions and Magnitudes, 283
Discrete Random Variables, 18

D...

Eigen Values and Eigen Vectors, 296
Estimation of the Regression Parameters, 155
Example of Half-Open Interval, 270
  - Open and Closed Set, 270
  - Compact Sets, 274
  - Complete Metric Space, 275

Existence and Uniqueness of Solutions, 151
Expectation of the Product of the Random Variables, 27
 – Sum of the Random Variables, 26

F...

Factorial Experiment, 192
Factorial Experiments with Factors at Two Levels (2^2 Factorial Experiment), 194
F-Distribution (Sampling Distribution of the Ratio of Two Sample Variances), 59
Few Examples of Connected Metric Spaces, 277
Field, 286
F-Test for Comparing Two Population Variances, 131

G...

Gauss-Markov Theorem, 173
Geometric Interpretation of the Solution set, 143
Geometry of Least Squares, 174
Graphical Representation, 11

H...

Hyperplanes and lines, 281
Hypothesis Test about a Variance, 110
 – Population Mean (μ) when the Population Standard Deviation (σ) is Unknown, 100
Hypothesis Test about the Population Mean (μ) when the Population Standard Deviation (σ) is Known, 91
 – Population Proportion (p), 106
Hypothesis Testing, 259, 262

I...

Importance of Compact Sets, 274
 – Metric Space, 269
Inclusion of Irrelevant Variables, 242
Independent and Dependent Random Variables, 26
Inferences about the Differences of Two Populations, 112
 – Two Means with
Independent Samples (Assuming Unequal Variances), 112
Inferences about Two Means with Dependent Samples–Matched Pairs, 120
 – Population Proportions, 126

J...

Joint Continuous Distribution Function, 25
 – Discrete Distribution Function, 23
 – Distribution Function and its Properties, 22
Jointly Distributed Random Variables, 21

K...

Key Properties of Compact Sets, 273

L...

Least-Squares Estimator, 166
Leverage and Influence Diagnostics, 234
Linear Dependence and Independence, 291
 – Equations, 136
 – Predictor and Link Functions, 257
 – Predictor and link functions, 261
Logistic Regression Model, 255

M...

Mallow's C_p Statistics, 246
Maximum Likelihood Estimation of Parameters, 258
Maximum Likelihood Estimation of Parameters, 262
Meaning of Regression Parameter, 166, 156
Measures of Centre, 4
 – Dispersion, 6
 – Influence, 236
Methods for Scaling Residuals, 214
Model Building, 244
Multiple Linear Regression Model, 164

N...

Normal Distribution, 42
  – Probability Plot, 218
Norms and Metric, 267

O...

Omission/Exclusion of Relevant Variables, 240
One-sample Test for Testing the Hypotheses, 111
One-Way Model with One Covariate, 206
Open set and Closed Set, 269
Open Set, 269
Orthogonal Projections, 176

P...

Partial F-statistics, 249
Pie Charts, 11
Plots of Residuals Against Explanatory Variable, 225
Plots of Residuals Against the Fitted Value, 222
Point Estimates, 75
Poisson Distribution, 37
  – Process, 40
Poisson Regression Model, 260
Pooled Two-sampled *t*-test (Assuming Equal Variances), 118
Power of the Test, 90
Press Statistic, 249
Probability Distribution, 29
Problems on Connected Metric Space, 277
Properties of the Estimated Parameters, 170
Properties of the Least Squares Estimators, 158
P-value Approach, 105, 117, 123

R...

Real analysis and Data Science, 264
Reduced row echelon form (RREF), 147
Residual Analysis, 213
Residual Mean Square, 245
Row Reduction and Echelon Forms, 144

S...

Sampling Distribution of Mean σ Unknown: *t*-distribution, 55
Sampling Distribution of Proportions 54
Sampling Distribution of Variance, 58
Sampling Distributions of Differences and Sums, 54
Sampling Distributions, 51
Sequences of Numbers, 271
Sequences, 271
Sequential Compactness, 273
Simple Linear Regression Model, 153
Single Factor ANOVA Model, 180
Solving Linear Systems, 140
Some Common Terms used in Sampling Distributions, 49
Some other Properties of Eigen values, 301
Spanning Sets, 287
Statistically Significant, 89
Stepwise Regression Techniques, 251
Subspace Formulation of Linear Models, 175
Subspaces, 288
Summary, 133
Sums of Squares, 185

T...

Testing a Hypothesis Using Classical Approach, 91
Testing a Hypothesis Using Classical Approach, 101
  – *P*-values, 96
Testing of Hypotheses (Overall F-test for the Equality of the Factor Level Means), 185
The Analysis of Covariance (ANCOVA) Model, 206
The Fundamentals of Hypothesis Testing, 85
The Linear Model for Two-way Factorial Experiments, 197

The Null and Alternative Hypotheses, 87
The Sampling Distribution of the Mean (σ known), 52
The Two-way Factorial Analysis, 197
Theorems Based on Connectedness of Metric Space, 276
Transformation and Weighing to Correct Model Inadequacies, 225
Transformations to Linearize the Model, 227
Two-Factor ANOVA Model, 188
Types of Errors, 89
 – Random Variables, 17

U...

Unbiases,171
Unions of Open Sets, Intersections of Closed Sets, 270
Use all Possible Explanatory Variables, 251
Using the *P*-Value Approach, 129

V...

Variable Selection, 239
Variance Stabilizing Transformations, 225
Vector Addition and Scaler Multiplication, 280
Vector Space, 287
Vector, 279

W...

What is a System of Linear Equations, 136
χ^2 – Distribution, 56

□□□

Win Prizes

ATTENTION READERS

Win Prizes

We request for your frank assessment regarding various aspects of the book as given below:

STATISTICS, STATISTICAL MODELLING AND DATA ANALYTICS

By: Rajni, Ajay Kumar Gupta, Kanika Agarwal

Please Fill up the Sheet in Neat and Clean Writing

(i) What topics of your syllabus, important for your examination point of view, are not covered in this book ?

....

....

....

....

(ii) What are the chapters wherein the treatment of the subject matter is not systematic or organised or updated?

....

....

....

....

(iii) Have you come across misprints/mistakes/factual inaccuracies in the book? Please specify the chapters and the page numbers.

....

....

....

....

(iv) What other books on the same subject have you found/heard better than the present book? Please specify in terms of price and quality

....

....

....

....

(v) Further suggestions and comments for the improvement of this book.

....

....

....

....

(PS: If need be please attach a separate sheet to write your views)

OTHER DETAILS

(i) Who recommended you this book? (Please tick (✓) in the box near the option relevant to you.

Teacher ☐ ☐ Friends ☐ Bookseller

....

....

....

....

(ii) Name of the recommending teacher, his designation and address:

....

....

....

....

(iii) Name and address of the bookseller you purchased the book from :

....

....

....

....

(iv) Name and address of your institution (Please mention the University or Board as the case may be)

....

....

....

....

(v) Your name and complete postal address:

....

....

....

....

(vi) Course you are enrolled in:

....

....

....

....

The best assessment will be awarded each month: The award will be in the form of our publications as decided by the Editorial Board.

Please mail the filled up coupon at your earliest to:

S.K. KATARIA & SONS®
4885/109, Prakash Mahal, Dr. Subhash Bhargav Lane,
Opposite Delhi Medical Association, Daryaganj, New Delhi–110002 (INDIA)
Phone: +91-11-23243489, 43551243; *Mobile:* +91-9871775858
e-mail: katariabook@gmail.com; katariabooks@yahoo.com
Website: www.skkatariaandsons.com;